Data Education for Children

Shengwen Dai

Master of Science

School of Informatics

University of Edinburgh

2019

Abstract

With the development of data science, it is becoming increasingly important for young people to learn this skillset. To explore this idea, our project adopts POGIL (process-oriented guided inquiry learning) to the field of data science education. We designed teaching materials of data science for young people following a POGIL approach. Our aim is to guide the students to understand the first four steps of the full cycle of data science workflow: data collection, data entry, data preprocessing, and data visualisation. The teaching material was evaluated twice, by a high school computer science teacher and two 16-year-old students respectively. The teaching material has generally achieved satisfactory results, with some areas worth improving.

Acknowledgements

Time crept on and one year passed. A year ago, I was a bachelor in mechanical engineering who could only write a few lines of code. A year later, I am about to become a master in artificial intelligence who can implement neural network algorithms. After completing the dissertation, I also mastered some practical teaching methods and understood how important it is to pass on what I have learned to the next generation.

Thanks to the professors and the tutors at the University of Edinburgh. You are so patient when answering my questions even though some of them are “stupid”.

Alone we can do so little; together we can do so much. Many thanks to all the teammates who have worked with me on the projects during this year: Hao, Zihan, Wenjun, Qiqi, Ying, Jiajun.

I cannot call it an easy year, but I really appreciate the happy time I spent with my good friends: Hao, Ziying, Xiaochu, Ruochen, Lin, Jack, Callum, Magali, Wei, Linlin. And many thanks to you for comforting me and guiding me when I was upset.

Wish we all realize our dreams and have ideal lives. Wish there is no longer disputes and wars in the world and people can live in an environment full of peace and love.

Table of Contents

Chapter 1

Introduction

In 2012, the Harvard Business Review described data science as the sexiest job of the 21st Century [37]. Data science combines knowledge from multiple disciplines including mathematics, statistics, computer science, and information science, using scientific methods, processes, algorithms and systems to analyze existing data [20, 12]. Currently, data science has evolved to include many additional aspects, such as business analytics [27], business intelligence, predictive modelling, and statistics.

Modern enterprises already make many decisions by analysing existing data. An important reason for this shift is that companies have become digital, with some companies primarily doing business through the Internet. Also, as individuals, communicating and socializing through the Internet has become an indispensable lifestyle. With the continuous development of computer science and the continuous acceleration of the Internet speed, data analysis work can be automatically completed with the support of new technologies. These factors make data-driven techniques more appealing and efficient [21, 22].

The interest in data science is growing rapidly. Many people believe that data science will be a competitive profession in the future. While computer science developed into an independent discipline in the 1970s, now we witness countries similarly accelerating the establishment of research centres for data science and the development of bachelor/master programmes in data science. Data is increasingly essential to both individuals and organizations and is becoming an indispensable resource.

Our overall aim is to develop a set of interactive teaching materials which convey key techniques in data science in an interesting and approachable way for young people from 10 to 14 years old. We propose three basic principles for teaching young students data science: stimulating creativity, stimulating interest, and emphasizing quality. We

designed teaching materials that followed the POGIL way.

Our project is an exploratory qualitative work. The evaluation of the teaching material is primarily based on feedback collected from experts (teachers) and users (children). Our goal was to teach children data literacy, enabling them to read and generate visualized representations of data while cultivating an interest in data science through our teaching material [10]. There are many interactive questions in our teaching material. Our research question is how to set the appropriate questions for students to get curious about the concept of data science.

In the second chapter, we introduce the background of data science discipline, which explores why students should learn data science and the current state of data science education in the world. The methodology is introduced in the third chapter, which details our teaching principles, teaching techniques, teaching plan, and summarizes the content of the teaching material. Evaluation sessions by experts and students are provided in the fourth chapter. Our findings after the completion of the project and the future work are presented in the fifth chapter. Conclusions are in the sixth chapter.

Chapter 2

Background

2.1 The development of Data Science

The modern definition of data science was first introduced at the University of Montpellier II (France) in 1992. Scholars participating in the second Japanese-French statistics symposium acknowledged that data science was a new discipline. These scholars shaped the contour of data science, based on the concepts of statistics and data analysis with the increasingly powerful computing power of computer tools [14]. In 1996, members of the International Federation of Classification Societies (IFCS) included "data science" in the title of the biennial conference held in Kobe. This was the first time the term "data science" was included in the title of a conference [28].

In April 2002, the first Data Science Journal was founded by the Committee on Data for Science and Technology (CODATA) of the International Council for Science (ICSU). This publication helped build knowledge and understanding of how data practices could advance research and human knowledge [5].In 2005, data scientists were defined as "the information and computer scientists, database and software and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection", by the publication "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century" published by National Science Board [31].

In 2012, Harvard Business Review published the article "Data Scientist: The Sexiest Job of the 21st Century" [37]. In this article, D.J. Patil and Jeff Hammerbacher coined the title "data scientist". The respective leads of data and analytics efforts at LinkedIn and Facebook then began to use this title to describe their jobs. In 2013, the IEEE Task Force on Data Science and Advanced Analytics (TF-DSAA) was launched.

In 2014, the IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA) was launched, in partnership with TF-DSAA [6]. In 2015, Springer started International Journal on Data Science and Analytics [7] to bring together thought leaders, researchers, industry practitioners, and potential users of data science and analytics.

2.2 The Necessity for Young People to Learn Data Science

In the past century, an essential factor of establishing an industrialized country has been to include compulsory subjects in elementary and secondary schools, such as mathematics, physics, chemistry and biology. The reason for this is not to train every child to become a scientist, but rather because the government recognizes that the development of society requires every citizen to understand some basic scientific concepts. Without mathematics, there is no economics; without physics, there is no engineering; without biology, there is no medicine.

Today's data science is like the basic science subject of a hundred years ago. This emerging discipline is growing rapidly. With the significant impact of the new science paradigm, educational programs in journalism, economic management, business, publishing science, biological sciences and social science also become more data-centric, prioritising their students awareness of and ability to tackle big data problems [38]. Being familiar with data science is as critical to every citizen as being familiar with traditional scientific disciplines. In order to be competitive in the fierce technological competition between a global tech economy in the 21st century, most countries are establishing thorough learning systems for young people to study digital literacy and informatics [17].

Data science has several good reasons to be included in the course system of young people. As a branch of informatics, data science focuses on using mathematical methods and informatics techniques to explore the information behind data. Thus, from the improvements that students can get while learning informatics, we can speculate on the abilities students can gain while learning data science. As [17] mentioned, informatics is especially essential in the modern education system, because learning data science can help students develop essential problem-solving skills. As a subject in the curriculum:

- Informatics promotes **creativity**: students can experiment with multiple ways to solve problems.

- Informatics is **constructive**: students go through a series of logical steps when designing algorithms to analyse data.

- informatics helps master **complexity**: students use the thinking skills acquired in the process of learning informatics when solving complicated problems in other fields.

Informatics focuses on how to use computers to do things that people can or cannot do, such as using algorithms to plan the most efficient travel routes; data science focuses on using algorithms to process real-life data, such as analysing a company's annual income. However, today, the boundary between the two fields is blurring and many of the problems associated with informatics need to be solved with the help of big data. Therefore, learners will also exercise the above three abilities when learning data science. These abilities can be taught, and must be taught, in primary and secondary schools.

In addition to the three abilities above, computational thinking, a problem solving process including logically ordering, analysing data, and creating solutions, is an essential part of data science work [3]. In [9], Valerie Barr and Chris Stephenson mention the importance of exposing students to computational thinking in the basic education stage:

*"It is no longer sufficient to wait until students are in **college** to introduce these concepts. All of todays students will go on to live a life heavily influenced by computing, and many will work in fields that involve or are influenced by computing. They must begin to work with algorithmic problem solving and computational methods and tools in K-12. "*

The restriction of "no longer sufficient to wait until students are in **college**" correlates with our following views:

- Not all young people continue their college education after basic secondary education. For the benefit of a countrys future informatization development, students who stop at the secondary school level will need to understand how to use modern technology to analyse data, just as they currently use mathematics to solve problems in daily life.

- Many students, regardless of their level of literacy, are likely to understand some data science technologies through informal channels. For example, some students count their actual expenses and make financial plans, or even use spreadsheets to process personal finance data. In general, their approach to data science technology is unstructured and non-systematic. But this way of learning does not allow young people to understand the basic principles of data science work. The role of data science education is to correct this erroneous learning method so that students can learn data science in a structured learning process based on their own creativity.

- Many university curriculums today require students to have necessary data analysis skills. This means that university teachers assume the majority of entering students are able to understand data science to some extent when designing their courses. Therefore, students should undergo formal training in data science before entering university.

- Beyond the specific ability to process data, almost all university curriculums require students to have some analytical skills. Data science education in the basic education stage thus can contribute a lot to developing students' problem-analysing ability.

From the analysis above, we conclude that data science education has two main roles in primary and secondary education: practical and educational.

- **Practical:** data science plays an essential role when students are looking for informatics-intensive jobs. Additionally, the analytical skills gained through learning data science will significantly help them succeed at their jobs.

- **Educational:** data science teaches students to look at problems in an objective way, meaning that students learn how to make rational decisions from a scientific point of view when approaching a new problem.

2.3 The State of Data Science Education

Data science is currently a top priority in the information society and an essential part of a country's industrial base. The degree of development of data science determines the innovation capability and technology level of a country to a certain extent. Data science develops differently in higher education and basic education.

Data science is a relatively new course. With the rapid development of informatization, today's universities have joined the data science curriculum with many other majors. The department or schools that provide data science courses include, but are not limited to, Business, Infomatics, Statistics, Mathematics, etc [36].

The widespread phenomenon of data science programmes in various universities illustrates the multidisciplinary nature of data science. [33] explored 48 data science projects in the United States and found that there are more master's programmes than bachelor's and doctoral programs. Most of these programs are joint projects developed by different departments or schools. [8] and [36] did the same kind of research and reached similar conclusions. They speculated that many universities were considering incorporating data science into undergraduate programmes.

However, the fast development of this discipline does not mean that data science education has found its place in the basic curriculum system. In fact, informatics education has retreated in most European curricula since pioneering efforts in the 1970s and 1980s [17].

Reports by the Association for Computing Machinery (ACM) and the Computer Science Teachers Association (CSTA) show that primary and secondary schools in Europe have made some progress in digital literacy, but still have a great room to improve data science education.

Since the 1970s some European countries have introduced informatics into primary and secondary education. These efforts have gradually dropped, because people do not correctly recognize the importance of informatics and lack sufficient experience to understand the importance of teaching digital awareness. As a branch of informatics, the development of data science education has also stagnated in recent years.

Chapter 3

Mathedology

What we are doing is a pioneering work - developing teaching materials regarding data science for young people who would have had little exposure to this discipline. In order for the data science curriculum to fulfill its practical and educational goal, we need to identify the basic principles of teaching data science and determine the best teaching method based on these principles before designing the teaching material itself. Additionally, we need to determine what knowledge can be accepted by students, as some concepts in this field are too complicated for young people to understand fully. The designed teaching materials are based on the rules and restrictions above.

3.1 Teaching Principles

Based on the current position of data science in the middle school curriculum system and the development of data science education, we consider three core principles when designing the curriculum: **stimulating creativity**, **stimulating interest**, and **emphasizing quality**.

3.1.1 Stimulating Creativity.

Certain characteristics of data science determine that it can significantly enhance students' creativity. First, innovation in the data science field is less difficult than in other disciplines. Second, with the development of science and technology, the technical equipment for data science research is getting cheaper. Therefore, data science sets a shallow barrier for beginners, allowing talented students to quickly learn the basic principles with proper guidance. Data science education workers should be aware of this

phenomenon and correctly guide students to apply creativity to do useful things. For example, teachers can show students some high-quality projects based on data science. Humanitarian Free and Open Software Systems (HFOSS) [11] tells how to apply data science to benefit society. Educators should also let students be aware of the ethical implications of their work.

3.1.2 Stimulating Interest.

Data science education should not only focus on teaching students the principle of analyzing data but should also pay attention to stimulating students' interest in data science work. For example, teachers can combine abstract concepts with real life problem-solving through visualization or animation.

3.1.3 Emphasizing Quality.

Some creative students learn data science by seeking resources around them, especially informal alternatives outside of formal data science education. The official data science education must emphasize quality, including hardware quality, software quality, textbook quality, and teacher strength. Teachers should also pay attention to their students' psychology and social factor while developing the data science education system.

From the perspectives above, data science education again has both **practical** and **educational** roles:

- **Practical:** Allows students to apply theory to practice. For example, it let students experience the difference between ”doing data science work” and ”processing data”. The former is a series of constructive activities based on science and engineering; the latter is only a specific step of the former.

- **Educational:** Lets students fully understand the importance of data science work and eliminate prejudice against this technology the idea that data or informatics are nerdy or unpopular fields. Emphasizing the people-oriented and user-centred nature of data science may also attract both genders to participate.

3.2 Teaching Technique

Since data science is a new discipline, there is still a lack of widely implemented lesson or teaching materials specially designed for young people. Educators are still exploring the most appropriate teaching methods for data science education. Data science workers need the ability to use technical tools for high-performance computing, as well as the process skills that help complete engineering processes. Thus, here we introduce the educational methods that have been applied in the computer science field and engineering field: process oriented guided inquiry learning.

3.2.1 Process Oriented Guided Inquiry Learning

Process oriented guided inquiry learning (POGIL) is a teaching method based on learning science (e.g. [39]). It is a student-centred, research-based pedagogic strategy that has been used originally in chemistry classrooms at all levels in colleges and high schools [24]. POGIL combines the characteristics of mainstream self-learning (e.g. active learning, discovery learning, and inquiry-based learning) and incorporates the idea of collaborative learning.

In POGIL sections, groups of learners (typically 3-5 people) conduct inquiry activities based on the designed teaching material and learn knowledge during this process. The learners in the team are assigned specific roles with different responsibilities and jointly promote the team to complete the learning steps and goals. Assigning roles can encourage learners to develop process skills and take responsibility for their work. The teacher's role is no longer that of a lecturer but rather that of an active facilitator. Thus, a very important point of the POGIL method is that students are actively learning knowledge rather than passively listening. POGIL activities and processes are designed to achieve specific learning objectives, such as learning certain academic concepts or engineering processes. Some workshops are presented in POGIL [4] to help teachers understand the basic principles and practical examples of POGIL, including how to develop, evaluate, and improve POGIL sections.

There are generally three learning phases in POGIL sections: **exploration**, **concept invention**, and **application**. The characteristics and functions of the three phases are as follows:

- **Exploration**: students infer certain trends or patterns from the data already provided or the data they collect, and validate the assumed concepts through subse-

quent questions.

- **Concept invention**: Concept invention: students learn new concepts through a hypothesises or concept obtained in the previous phase. Unlike a traditional class, students have already constructed a certain understanding of the new concepts through inquiry learning before introduced the concepts.

- **Application**: students do practical activities with the learned concepts. In this process, students appreciate the meaning and applicability of the new concept.

Therefore, the scripted activity above provides students with the necessary materials and questions to learn the concepts and guide students through the learning cycle to help them gain academic knowledge and process skills.

In POGIL sections, the questions that need students to inquire actively are mainly divided into three categories: **directed question**, **convergent question**, and **divergent question**. Their characteristics are as follows:

- **Directed question**: There is a clear answer. The question is based on the student's previous knowledge or provided material. The answer to the question is the basis for the follow-up activity.

- **Convergent question**: There might be multiple answers. The answer to the question is not an obvious conclusion and needs to be discussed by the team members.

- **Divergent question**: Groups may move in different directions depending on which answers they come up with.

3.2.2 Applying POGIL to Design the Teaching Material

We designed the teaching material following the POGIL method. Applying POGIL to the data science field was not an easy task. At present, there are not many courses on data science designed with POGIL, and the development of data science is faster than the development of subjects like chemistry, and some content is updated faster. Thus, educators need to put significant effort into developing the POGIL teaching material regarding data science.

Although teaching data science in the form of POGIL is costly and time-consuming for teachers, its benefits for students far exceed traditional methods. For data science

education, POGIL has special potential. Generally, exploring a full cycle of data science is a relatively suitable course theme but when students use POGIL to learn data science, they gradually gain problem-solving skills they can apply right away. Teamwork is a very important skillset for completing future tasks that students will face, and POGIL can help students improve their ability to communicate and cooperate with peers. Additionally, POGIL emphasises interaction between students, rather than just listening to a lecturer. Because of the different characteristics of each discipline, POGIL, a teaching method originating from chemistry, cannot be applied by other disciplines without modification. If used properly, the benefits that POGIL brings to students are far greater than the negatives.

POGIL fosters independent thinking and allowing students to explore concepts on their own. The focus on interaction and communication also encourages students to work in teams. POGIL is by far the best method we found to teach young people data science. We carried out our projects in POGIL way.

3.3 Teaching Plan

3.3.1 Teaching aim

Some papers discussed the motivations for teaching young people about data science. We classify these motivations into two categories: **learning the actual skills** (e.g. [13]) and **being engaged in data science** (e.g. [32]).

- **Learning the actual skills**: help students master the concepts of data science and implement the full cycle of data science completely. Doing so allows students to master computational thinking and use this way of thinking in any field.
- **Make students interested in data science**: let students have a sense of excitement while learning data science. Help students master some of the concepts of data science and the engineering process (like mastering basic mathematics in primary and secondary education). Skill development is the second concern.

At present, data science is not considered a compulsory course in primary and secondary schools. It is unrealistic for all students to master the data science workflow, so our purpose is similar to the second one. We hope that our courses allow young people to abandon the negative impressions of data science and make them feel that the world of data is interesting.

3.3.2 Teaching content

In modern data science, a full cycle of data science work is composed of data collection, data entry, data preprocessing, data visualisation, model building, model testing [35]. The last two steps in the full cycle are generally used for supervised learning and unsupervised learning. For most young people, understanding the algorithms of machine learning is somewhat tricky. In our teaching plan, model building and model testing are not discussed in detail. In order to stimulate students' interest in data science, the content in our teaching materials should not be far from the students' general knowledge (for example, students have not been exposed to stocks, thus introducing stocks in the teaching material are not suitable). The examples in the teaching material should be closely **relevant** to students' daily life. We will discuss the "**relevant**" in detail in the next section.

Our teaching material guides students to experience the first four steps of data science work - data collection, data entry, data preprocessing, data visualisation. In order to enhance the practicality of data science in students' daily life, we also introduce how to read infographic in the teaching material.

3.4 Helping Students Find Relevance

As we mentioned in the last section, we hope that our courses allow young people to abandon their negative impressions of data science and make them feel that the world of data is interesting. To achieve our aim, one of the most important points is our teaching material is designed to enhance the learners engagement and motivation in data science.

3.4.1 Relevance: the Core Element

In the teenage years, when the teacher was giving lectures, students might wonder, "yeah, but what am I gonna use this for?" or "what does this have to do with me?" These kinds of questions usually appear in compulsory courses (our electives are decided by ourselves, so these questions were unlikely to appear), because we cannot find content that is worth our time or energy in the classroom. When students ask such questions, it is not because the content of the lesson is not important, but because they are looking for relevant links between themselves and the curriculum.

Relevance is a difficult concept to pin down. It is mentioned in the education

literature, but usually as an aside and seldom with an explanation as to its nature or structure [29]. In [25], Relevance is defined as a perception that is triggered when someone thinks something is **interesting** and **worth knowing**. In [34], Wilson and Sperber proposed this theory during the 1980s:

"...utterances raise expectations of relevance not because speakers are expected to obey a Co-operative Principle and maxims or some other specifically communicative convention, but because the search for relevance is a basic feature of human cognition, which communicators may exploit."

This theory may sound somewhat Machiavellian, but it provides a theoretical basis for us to communicate our intentions to students: by tapping into the cognitive world of students, we can better communicate our intentions and make the work more relevant to students. It is important for teachers to provide relevance to students because it is related to mobilising students' interest in learning new knowledge [15, 23].

Back to the definition, **relevance is the perception that people get when they think about something interesting and worth knowing.** It is worth noting that there are two main points in the definition: **interesting** and **worth knowing**. Many teachers make the lesson more interesting by adding "interesting things" to the class and expecting to increase the relevance between students and the class topic. The "interesting things" added to the lesson plan include, but are not limited to, jokes, games, etc. These "interesting things" may attract students' attention at the beginning, but it is not easy to make students stay attentive throughout the class. Once the "boring" knowledge is introduced, the student's attention will be diminished if the knowledge is not explained in a way that the students like or can relate to. Here, we have not denied the importance of incorporating "interesting things" into the classroom; rather, we emphasise that these "interesting things" need to be relevant to "boring things" in order to keep students motivated during the whole lesson.

3.4.2 Two Basic Ways to Provide Relevance for Students

Roberson [29] introduced two basic ways to provide relevance for students: **utility value** and **relatedness**.

3.4.2.1 Utility Value

Utility value can help us solve one of the problems mentioned above, "Yeah, but what am I gonna use this for?". Utility value is purely academic, which shows students

the importance of what they are currently learning for their future career - both short-term and long-term goals [25]. For example, physics is boring for ordinary students, but for students who want to be scientists or engineers, physics is interesting and has a lot of utility value. There are two main steps to using the utility value to show relevance for students: first, pique the students' interest by telling them that what they are learning will be important for their future career goals, then explaining why what they are learning is worth knowing, and subsequently how the knowledge they are learning helps them achieve future career goals. This approach allows students to realise that the knowledge is not only interesting but also worth knowing.

3.4.2.2 Relatedness

Prioritising relevance answers the question "What does this have to do with me?". Relevance refers to an inherent need for humans to feel close to the significant people in their lives[30]. Many studies consider relevance to have both non-academic and academic applications. In terms of teaching activity, the non-academic side of close relevance is the relationship between the teacher and the student. Students are more likely to listen to the lesson when they feel closer to the teacher, as students are more likely to agree with people they respect. Students value the knowledge taught by their favourite teachers and trust that this knowledge as something worth learning because the teachers tell them that it is something worth knowing. This is why it is crucial to have the teacher show enthusiasm during the teaching activities. The academic aspect of relevance emphasises that teachers should help students connect with the knowledge they need to learn and develop their future careers. Students can perceive the teachers' efforts to relate to them and interpret them as a sign that their teachers care. Students respond to this care by putting effort into learning the knowledge taught by the teacher. The process of providing relevance for students by relating to them is also divided into two steps: first, teachers establish close relationships with students to stimulate students' interest in their lesson, then teachers connect that intellectual curiosity with the knowledge they need to develop their future careers, thus helping students understand the value of what they are learning.

3.4.3 Providing relevance in the Teaching Material

One of the most important things that educators need to do is to provide relevance for students. When students think that what they need to learn is relevant to their life,

they are more likely to become motivated learners. For a student, relevance is vital at all ages and its utility value increases as the student ages. Relatedness can play a significant role in the students' compulsory course, as it helps students realize that all knowledge is worth knowing. Utility value can help students determine which courses may be helpful for future career development, once students begin to choose optional courses. When we designed the data science teaching material, we provided relevance through relatedness and utility value. For example, we could link interesting examples to boring concepts, or combine nebulous knowledge with a students daily lives, or explain why the knowledge they were learning is helpful for their future career. In this way, students are more likely to believe that the content of our material is interesting and worth knowing, thus helping them become an active learner..

3.5 Designing the Teaching Material

According to the basic principles of POGIL, we divide the teaching material into three parts: exploration, concept invention, application. Each part requires learners to use knowledge more skillfully than the previous part. In this section we introduce the three parts of the teaching material.

In the exploration phase, we wanted students to understand what data science is and how data science works. During this phase, the teaching material focuses on data preprocessing and data visualisation. In the concept invention phase, students use the concept of data visualisation learned in the previous phase to explore three different visualisation techniques, and use one of them to present the data set used in the exploration phase. In the application phase, we designed two hands-on exercises to enhance students' data analysing capabilities and data understanding capabilities. Here, we introduced the dataset used in the teaching material in 3.5.1, and outline the descriptions of the three phases in the following sections. Please note that all the description about the teaching material in this section are the first version before the evaluation session. The improved version after the evaluation is introduced in the chapter 4.

3.5.1 The Coffee Bean dataset

As we mentioned in the previous subsection, the content should be relevant to students' daily life and so we knew we should present a dataset that most students would be interested in. Topics such as food, sports, and pets are raised at a high frequency among

young people. Moreover, these topics are the "safe topics" that we can incorporate into the teaching material, compared to sensitive topics such as politics and international disputes. Thus, we used the data set related to coffee beans as our sample data set. Half of the lesson is related to this dataset.

We divided coffee beans into two species, Arabica and Robusta. We selected 1340 reviews of coffee beans grown in 35 countries and regions. Each coffee bean was rated from 10 features such as aroma, flavour, aftertaste, and acidity. We added up these 10 single scores to get the quality score. The data set is from the Coffee Quality Institutes (CQI) trained reviewers [2] and was cleaned by Github user jldbc [1]. We show the original dataset and the activities we designed with this dataset in the following subsections.

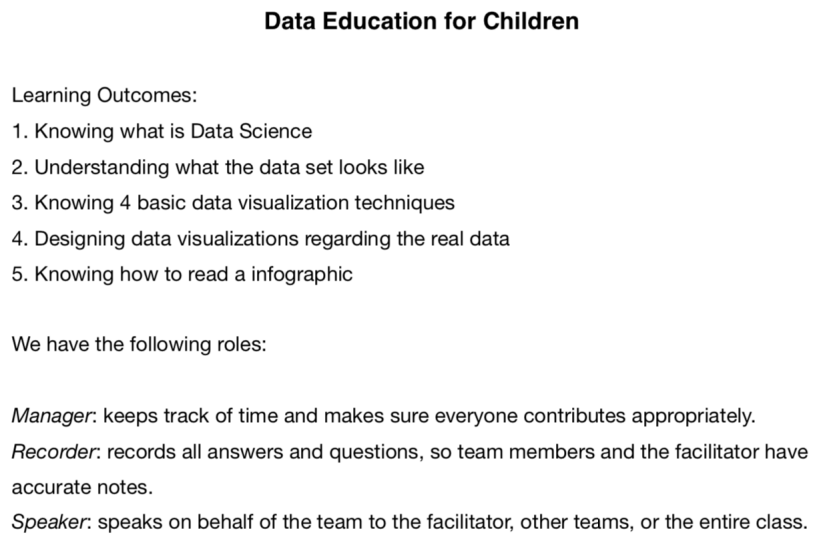
3.5.2 Exploration

Figure 3.1 shows the first page of the teaching material. At the beginning of the lesson, we need to inform students about the day's learning outcomes - letting students know what the specific knowledge for today is. Our teaching material is designed following the POGIL method, which emphasizes teamwork among students. Thus, we need to let students choose from the specific roles based on the requirements of the teaching material. The characters to choose from are *manager*, *recorder*, *speaker*.

- *Manager*: keeps track of time and makes sure everyone contributes appropriately.
- *Recorder*: records all answers and questions, so team members and the facilitator have accurate notes.
- *Speaker*: speaks on behalf of the team to the facilitator, other teams, or the entire class.

After the students understand the content of today's lesson and assign the roles, the exploration phase begins. According to the definition of exploration phase, *students infer certain trends or patterns from the data already provided or the data they collect, and validate the assumed concepts through subsequent questions*, we need to show students some reasoning materials (data) to let them think critically about certain trends or patterns.

We hope that the reasoning materials we designed can not only reflect some aspects of data science but also be relevant to the student's daily life - doing so can increase



Data Education for Children

Learning Outcomes:
1. Knowing what is Data Science
2. Understanding what the data set looks like
3. Knowing 4 basic data visualization techniques
4. Designing data visualizations regarding the real data
5. Knowing how to read a infographic

We have the following roles:

Manager: keeps track of time and makes sure everyone contributes appropriately.
Recorder: records all answers and questions, so team members and the facilitator have accurate notes.
Speaker: speaks on behalf of the team to the facilitator, other teams, or the entire class.

Figure 3.1: The first page of the teaching material

students interest. Thus, in the first activity we show three examples presented in the teaching material, the recommendation system of YouTube, the weather of Edinburgh, and the friends suggestion list of Facebook. After reading these three examples, students discuss the following question *"(3min) Question 1: use your words to conclude what Data Science is."*. Figure 3.2 shows two of these examples.

Students are then exposed to the first two data sets of the lesson. Information on 1312 Arabica and 28 Robusta coffee beans was recorded in two independent data sets, respectively. The information is the country (region) of origin, company name, planting altitude, quality score, etc. of each coffee bean. We guided students to compare the quality of each coffee bean by country of origin, based on quality scores. Figure 3.3 shows a portion of the data set of the Arabica coffee bean.

The current data set is too chaotic for the data analysis work that will be performed next. In order to make sense of it we need to preprocess the data - delete the unwanted data and reorder the features. Spending class time teaching students how to use software in order to do data processing is time-consuming and does not allow students to concentrate on the full cycle of data science. Therefore, we presented the cleaned data directly on the teaching material (figure 3.4).

-Exploration-

Let' see three examples first.

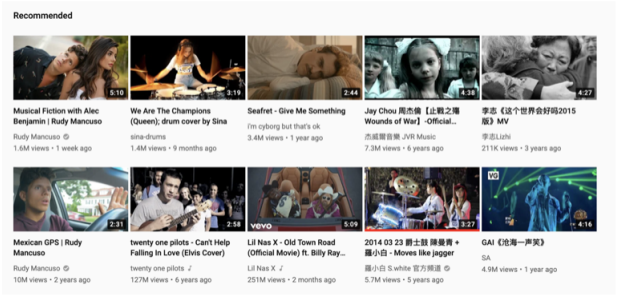


Figure 1: the recommendation system of Youtube

Figure 2: the weather forecast of Edinburgh.

Figure 3.2: two examples for students to infer what Data Science is.

	quality_score	Species	Owner	Country of Origin	Farm Name	Lot Number	Mill
0	90.58	Arabica	metad plc	Ethiopia	METAD PLC		METAD PLC
1	89.92	Arabica	metad plc	Ethiopia	METAD PLC		METAD PLC
2	89.75	Arabica	Grounds for Health Admin	Guatemala	San Marcos Barrancas "San Cristobal Cuch		
3	89	Arabica	Yidnekachew Dabessa	Ethiopia	Yidnekachew Dabessa Coffee Plantation		Wolensu
4	88.83	Arabica	metad plc	Ethiopia	METAD PLC		METAD PLC
5	88.83	Arabica	Ji-Ae Ahn	Brazil			
6	88.75	Arabica	Hugo Valdivia	Peru	n/a		HVC
7	88.67	Arabica	Ethiopia Commodity Exchange	Ethiopia	Aolme		C.P.W.E
8	88.42	Arabica	Ethiopia Commodity Exchange	Ethiopia	Aolme		C.P.W.E
9	88.25	Arabica	Diamond Enterprise Plc	Ethiopia	Tulla Coffee Farm		Tulla Coffee
10	88.08	Arabica	Mohammed Lalo	Ethiopia	Fahem Coffee Plantation		
11	87.92	Arabica	CQI Q Coffee Sample Representative	United States	El filo		
12	87.92	Arabica	CQI Q Coffee Sample Representative	United States	Los Cedros		
13	87.92	Arabica	Grounds for Health Admin	United States (Hawaii)	Arianna Farms		
14	87.83	Arabica	Ethiopia Commodity Exchange	Ethiopia	Aolme		C.P.W.E
15	87.58	Arabica	CQI Q Coffee Sample Representative	United States	El Águila		

Figure 3.3: A portion of the data set of the Arabica coffee bean.

	Species	Country.of.Origin	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Clean.Cup	Sweetness	Cupper.Points	Total.Cup.Points
1	Arabica	Ethiopia	8.67	8.83	8.67	8.75	8.5	8.42	10	10	10	8.75	90.58
2	Arabica	Ethiopia	8.75	8.67	8.5	8.58	8.42	8.42	10	10	10	8.58	89.92
3	Arabica	Guatemala	8.42	8.5	8.42	8.42	8.33	8.42	10	10	10	9.25	89.75
4	Arabica	Ethiopia	8.17	8.58	8.42	8.42	8.5	8.25	10	10	10	8.67	89
5	Arabica	Ethiopia	8.25	8.5	8.25	8.5	8.42	8.33	10	10	10	8.58	88.83
6	Arabica	Brazil	8.58	8.42	8.42	8.5	8.25	8.33	10	10	10	8.33	88.83
7	Arabica	Peru	8.42	8.5	8.33	8.5	8.25	8.25	10	10	10	8.5	88.75
8	Arabica	Ethiopia	8.25	8.33	8.5	8.42	8.33	8.5	10	10	9.33	9	88.67
9	Arabica	Ethiopia	8.67	8.67	8.58	8.42	8.33	8.42	9.33	10	9.33	8.67	88.42
10	Arabica	Ethiopia	8.08	8.58	8.5	8.5	7.67	8.42	10	10	10	8.5	88.25
11	Arabica	Ethiopia	8.17	8.67	8.25	8.5	7.75	8.17	10	10	10	8.58	88.08
12	Arabica	United States	8.25	8.42	8.17	8.33	8.08	8.17	10	10	10	8.5	87.92
13	Arabica	United States	8.08	8.67	8.33	8.42	8	8.08	10	10	10	8.33	87.92
14	Arabica	United States (Hawaii)	8.33	8.42	8.08	8.25	8.25	8	10	10	10	8.58	87.92
15	Arabica	Ethiopia	8.25	8.33	8.5	8.25	8.58	8.75	9.33	10	9.33	8.5	87.83

Figure 3.4: The preprocessed data set of Arabic coffee beans.

We took the first five rows of data from the preprocessed data set and rearranged their order to present the student with a new data set 3.5. Then we asked them the second question, *(2min) Question 2: compare the qualities of the first five coffee beans regarding the Total.Cup.Points in the table below.* This question works to address the third question, *(3 min) Question 3: is there a faster way other than just looking at the figures? Hint: recall the examples at the beginning of the class.* We hope to use these two questions in tandem to make students realise that the reasonable use of **data visualisation** can be more efficient when we compare data.

	Species	Country.of.Origin	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Clean.Cup	Sweetness	Cupper.Points	Total.Cup.Points
1	Arabica	Guatemala	8.42	8.5	8.42	8.42	8.33	8.42	10	10	10	9.25	89.75
2	Arabica	Ethiopia	8.75	8.67	8.5	8.58	8.42	8.42	10	10	10	8.58	89.92
3	Arabica	Ethiopia	8.25	8.5	8.25	8.5	8.42	8.33	10	10	10	8.58	88.83
4	Arabica	Ethiopia	8.67	8.83	8.67	8.75	8.5	8.42	10	10	10	8.75	90.58
5	Arabica	Ethiopia	8.17	8.58	8.42	8.42	8.5	8.25	10	10	10	8.67	89

Figure 3.5: The reordered data set.

The Arabica coffee beans dataset contains information on 1,312 coffee beans; these coffee beans come from 35 countries and regions. When we compare the quality of coffee beans by country/region, we need to use a statistical term to measure the scores of different countries/regions. We asked students the fourth question: *(3min) Question 4: we got 1312 Arabica coffee beans from 35 different countries and regions (the Table only shows 15 of them). What statistical term do we use to compare the quality of their coffee beans by country and region?*

(3 min) Question 4: we got 1312 Arabica coffee beans from 35 different countries and regions (Table 2 only shows 15 of them). What statistical term do we use to compare the quality of their coffee beans by country and region? Review status with the facilitator before continuing.

Figure 3.6: The fourth question and the breakpoint.

In the teaching material, the answer to this question is unique average score (of course, in real life we can also measure the quality of coffee beans with other statistical terms such as maxima or median). We added a note *"Review status with the facilitator before continuing."* after this question as we hoped to unify the progress of each group after this question, to give the slower groups a chance to catch up with the faster groups. We explained why the average score is the correct answer after all the teams answered the question, because the answer to this question will affect whether the group is on the right track when they move on to the next question. The fourth question and the breakpoint is shown in Figure 3.6.

So far, students have learned that data visualisation is efficient in comparing data and have learned how to use statistical methods to capture the hidden information behind the existing data. We then designed activities in the teaching material to let students experience the process of generating data visualisation. We calculated the average scores of coffee beans from Colombia, India, China, Uganda, United States, Ecuador, Ethiopia, Japan for students, as a preparation for the fifth question. In the fifth question, we asked students to use pencils to complete the unfinished data visualisation provided by us, *(3 min) Question 5: The average scores of the coffee beans from Colombia, India, China, Uganda, the United States, Ecuador, Ethiopia, Japan are provided below. Complete the bar chart. Review status with the facilitator before continuing.* There is again a breakpoint as we want to make sure that each group knows how to implement the bar chart. The table and the bar chart that needs to be supplemented by the students is shown in Figure 3.7.

Our data set was the quality score of 1312 Arabica and 28 robusta coffee beans from 35 countries and regions. As the lesson progresses, students are asked to preprocess the data and calculate the average score based on the country and region. Finally, they are asked to generate a bar chart to visually compare the scores of different countries and regions. Students completed the comparison between eight countries, and

found that the comparison between the 35 countries only required more work than the former, which means that the students would not learn new knowledge in this process. Thus, we showed students the complete table and data visualisation in 35 countries and regions on the teaching material (Figure 3.8).

	Country of origin	Average Score
1	Colombia	83.11
2	India	76.83
3	China	82.93
4	Uganda	84.05
5	United States	85.98
6	Ecuador	83.83
7	Ethiopia	84.54
8	Japan	84.54

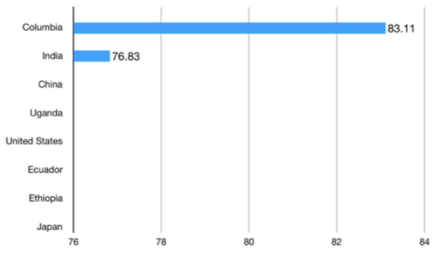


Figure 3.7: The table and the bar chart that needs to be supplemented by the students.

	Country of origin	Average Score (Arabica)
1	Colombia	83.11
2	India	76.83
3	China	82.93
4	Uganda	84.05
5	United States	85.98
6	Ecuador	83.83
7	Ethiopia	84.54
8	Japan	84.54
9	Vietnam	85.75
10	Kenya	84.31
11	Panama	83.71
12	El Salvador	83.05
13	Rwanda	82.83
14	Costa Rica	82.79
15	Thailand	82.57
16	Indonesia	82.57
17	Peru	82.53
18	Brazil	82.41
19	Tanzania	82.37
20	Zambia	81.92
21	Guatemala	81.85
22	Laos	81.83
23	Burundi	81.83
24	United States (Hawaii)	81.82
25	United States (Puerto Rico)	81.73
26	China (Taiwan)	82.00
27	Malawi	81.71
28	Mexico	80.89
29	Philipines	80.83
30	Myanmar	80.75
31	Mauritius	80.50
32	Nicaragua	80.46
33	Honduras	79.36
34	Haiti	77.18
35	Papua New Guinea	85.75

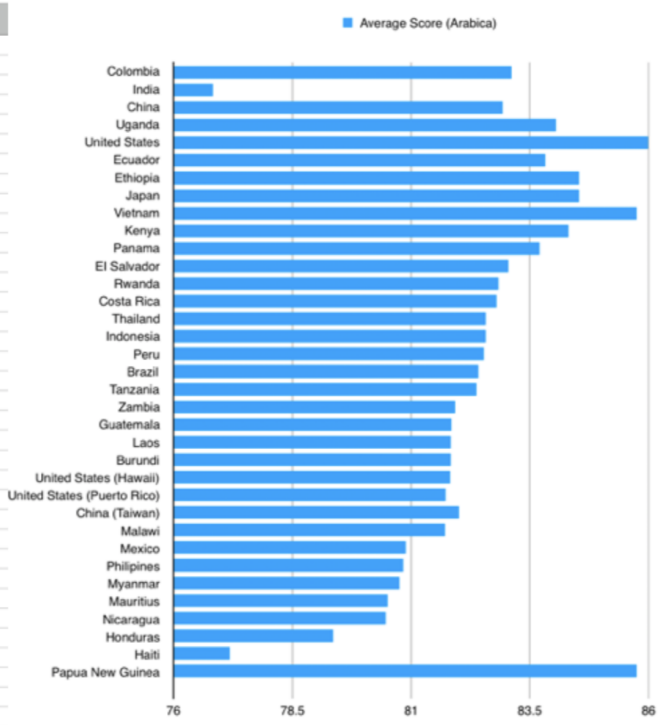


Figure 3.8: The complete table and data visualisation of Arabica coffee beans in 35 countries and regions.

Figure 3.8 shows only the data of Arabica coffee beans, but not the Robusta coffee

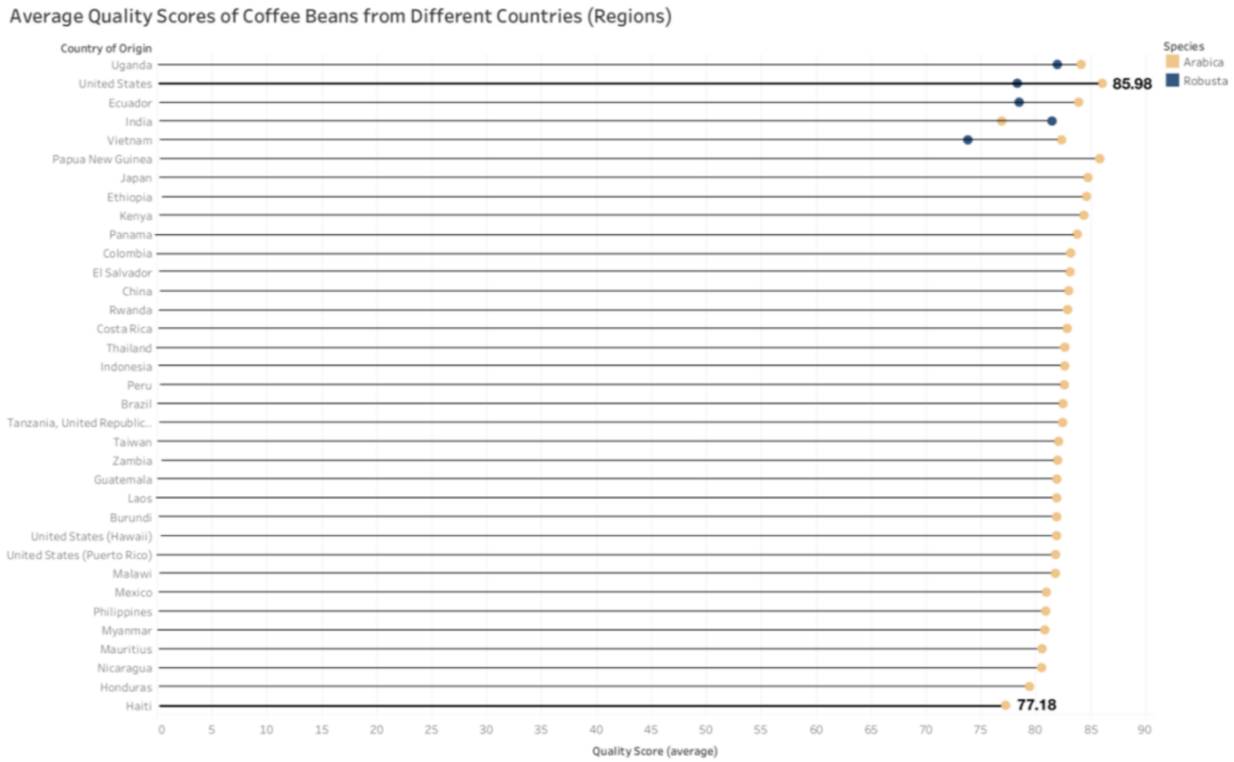


Figure 3.9: The complete data visualisation of Arabica and Robusta coffee beans in 35 countries and regions.

beans. In order to present students with a complete data science workflow, we need to explain to them that our work has not been completed yet, because the current data visualisation is incomplete. In the teaching material we mentioned: Do you remember there are two independent datasets? The visualisation above is for Arabica coffee beans. We still need to complement the quality score of Robusta coffee beans to our data visualisation. Similarly, we do not need students to complete this step. We present the generated data visualisation on the teaching material (Figure 3.9).

So far, we have finished the exploration phase. Students have learned that data visualisation is efficient in comparing data and can make use of statistical methods to capture the hidden information behind the existing data. Students also experienced the process of generating the complete data visualisation through group collaboration.

3.5.3 Concept Invention

Concept invention is a process whereby *students learn new concepts through the hypothesises or concepts obtained in the previous phase.* Unlike the traditional class, students have already constructed a certain understanding of the new concepts through inquiry learning before the teacher formally introduces the concepts.

In the concept invention phase, we hoped that students could actively learn the characteristics and the scope of application of the three visualisation techniques outlined in our reasoning materials. Before presenting the data visualisation techniques, we first set the time for the students to read the materials, *(12 min). Question 6: Read these data visualisation techniques.*

Pie Chart #1.

English dialestc	Proportion
American	70.7%
British	15.9%
Canadian	4.9%
Austrilian	4.8%
Other	3.7%

Table 6

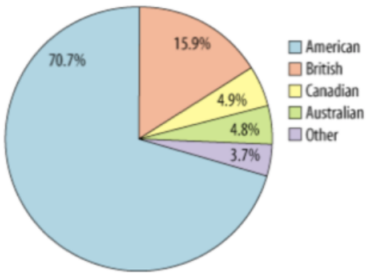


Figure 6

The table and chart show the relative numbers of native English speakers in the major English-speaking countries of the world.

Pie Chart #2.

	Purpose	proportion
1	Shower	16.8%
2	Toilet	26.7%
3	Leaks	13.7%
4	Faucet	15.7%
5	Clothes Washer	21.7%
6	Other	5.3%

Table 7

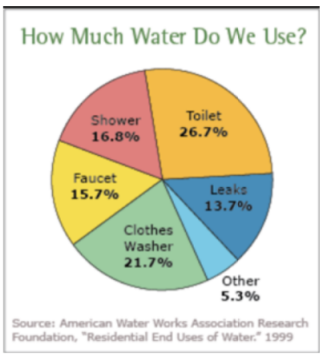


Figure 7

The table and chart show residential end uses of water in 1999.

Figure 3.10: The two examples of the pie chart.

The pie chart is widely used in our daily life; young people are familiar with the graphic in magazines and the news as a way to present information. Learning how to read and generate a pie chart is essential for students to communicate their findings. In the teaching material, we have prepared two examples which can be represented by the pie chart. The first example concerns the relative numbers of native English speakers in the major English-speaking countries of the world, and the second example involves the residential end uses of water in 1999. Each example is represented by a table and a chart (3.10). We hope that students can determine the characteristics of the pie chart by reading these two examples - the data set is composed of different items, with the value of each item expressed as a percentage, and all the values adding up to 100%.

Map Chart #1.

	Country	Phone Lines
1	Afghanistan	101931.0
2	Albania	235734.0
3	Algeria	3098787.0
4	American Samoa	9900.0
5	Andorra	38237.0
6	Angola	281327.0
7	Antigua and Barbuda	19915.0
8	Argentina	9602605.0
9	Armenia	572772.0
10	Aruba	35000.0
11	Australia	9190000.0
12	Austria	3254700.0
13	Azerbaijan	1795448.0
14	Bahamas	125658.0
15	Bahrain	284684.0
16	Bangladesh	974181.0

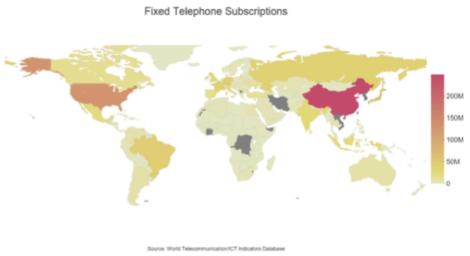


Table 8 **Figure 9**

The table and chart show the distribution of fixed telephone subscription in the world.

Map Chart #2.

Name	population
London	7556900
Birmingham	984333
Liverpool	864122
Nottingham	729977
Sheffield	685368
Bristol	617280
Glasgow	591620
Leicester	508916
Edinburgh	464990
Leeds	455123
Cardiff	447287
Manchester	395515
Stoke-on-Trent	372775

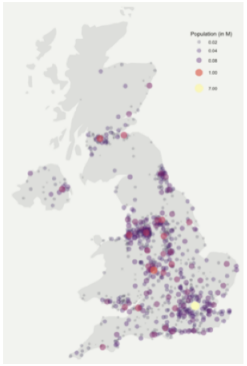


Table 9 **Figure 10**

The table and chart show the 1000 biggest cities and towns in the UK by population in

Figure 3.11: The two examples of the map chart.

The second visualisation technique in the teaching material is the map chart. The map chart marks different geographic locations on the map to show information about

each locations individual attributes. The map chart is widely used in many topics such as political, economic, and cultural. Map charts allow people to easily summarise and understand geographically relevant information. We prepared two examples which could be represented by the map chart. These two examples concern the distribution of fixed telephone subscriptions in the world and the 1000 biggest cities in the UK by population in 2019. Each example is represented by a table and a chart (3.11). We hope that students can reason that the data set represented by the map chart must be related to a geographic location.

Line Chart #1.

Time	Temperature
00:00, Sun 30 Jun	14
01:00, Sun 30 Jun	15
02:00, Sun 30 Jun	15
03:00, Sun 30 Jun	15
04:00, Sun 30 Jun	14
05:00, Sun 30 Jun	14
06:00, Sun 30 Jun	15
07:00, Sun 30 Jun	15
08:00, Sun 30 Jun	16
09:00, Sun 30 Jun	16
10:00, Sun 30 Jun	18
11:00, Sun 30 Jun	16
12:00, Sun 30 Jun	17
13:00, Sun 30 Jun	18

Table 10

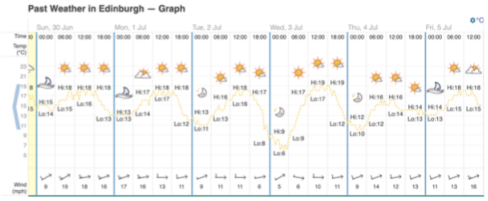


Figure 11

The table and chart show the weather of Edinburgh from 30/06 to 05/07.

Line Chart #2.

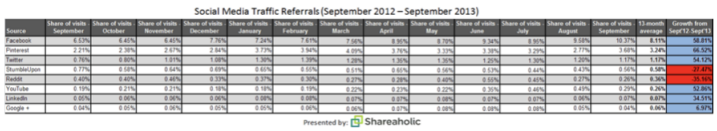


Table 11

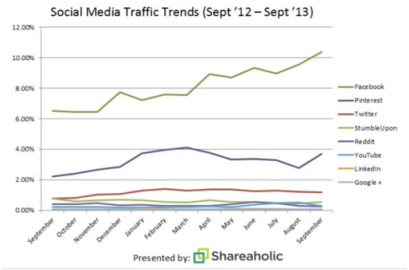


Figure 12

The table and chart show social media traffic trends (Sept'12 - Sept'13).

Figure 3.12: The two examples of the line chart.

The third visualisation technique in the teaching material is the line chart. A line chart is often used to visualize a trend in data over intervals of time - a time series - thus the line is often drawn chronologically. The line chart is ubiquitous in the topics like

economics, politics, and history. The two examples represented by line chart are the weather of Edinburgh from 30/06 to 05/07 and the social media traffic trends (Sept12 - Sept13).Each example is represented by a table and a chart 3.12.

After students have read these three examples, they will see the seventh question, *(4 min) Question 7: Find out the commonalities in each of the visualisation techniques. In what situations do we use a pie chart, map chart, and Line chart?* We hope that students can determine that each data set has an appropriate corresponding value: in the data set represented by the pie chart, all the values of the items are percentages and the percentages add up to 100%, while the data represented by the map chart concerns geographical locations, and the data set represented by the line chart tabulates time series.

So far, students have learned about the characteristics and scope of application of the three data visualisation techniques. Some data sets may be better represented by multiple visualisation methods different from the ones proposed by us. To exercise their divergent thinking and practical skills, we address the eighth question, *(5 min) Question 8: Based on the answer of question 7, what other data visualization can we use to compare the quality of coffee beans by different countries and regions? Why? Sketch the visualisation to verify your assumption. Review status with the facilitator before continuing*, to let students combine their newly learned concepts with the data sets we have used before. Here we set a breakpoint as we wanted to make sure that each group understood which was the most appropriate way to visualise data beyond the stated visualisation methods presented by us. After each group understood the reason why we used a map chart here and how to sketch a map chart, we provided the map chart in the teaching material (Figure 3.13). Then we taught students about how to read this map chart through oral instruction:

- The small and big bubbles represent the countries and regions where the Arabica and Robusta coffee beans are planted, respectively.
- The brightness of bubbles indicates the average quality score of a specific country or regions coffee beans. Uganda, the US, Ecuador, India, and Vietnam planted two species of coffee beans.
- The area where coffee beans are planted are mainly distributed in eastern Africa, southeast Asia, and central America. These areas are all around the equator.

So far, students have understood the characteristics and the scope of application of the

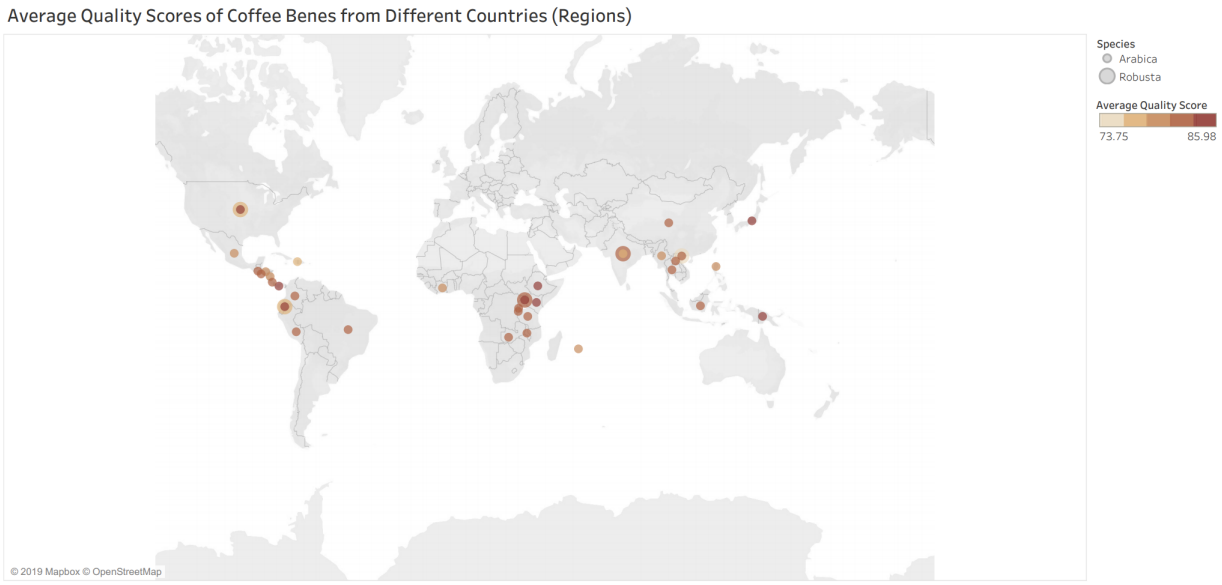


Figure 3.13: Average quality scores of coffee beans from different countries and regions.

three data visualisation techniques. They have also experienced divergent thinking by considering alternative data visualisations to express the same data set.

3.5.4 Application

Our teaching material had two goals. The first was to lead students to experience the first four steps of a data science workflow - data collection, data entry, data preprocessing, and data visualisation. The second goal was to help students develop the ability to read an infographic. To achieve these two goals, we designed two hands-on exercises in the teaching material.

3.5.4.1 M&M Peanut Chocolates Exercise

Up to this point students experimented with data preprocess and data visualization by using the dataset provided by us. Next, we asked each group to complete the following four steps, which meant that they needed to collect data by themselves.

In this hands-on exercise, the props we used were M&M peanut chocolates. We distributed the peanut chocolates to the students and asked the ninth question in the teaching material, *(2 min) Question 9: Open the M&M's peanut chocolate and taste some of them* turning the props themselves into snacks.

Our next question was: *(10 min) Question 10: In the remaining chocolates, select one or more features (ask the facilitator if you are not sure about what a 'feature' is). Then generate your dataset (in the table form) and express some of the features in a visualisation way that you think is appropriate.* These questions would lead the students to complete the following four steps in the data science workflow. First, we asked the students to think about the features of each chocolate and collect the data according to these features; second, students filled in the form with the collected data (data entry); third, students chose the features (data preprocess) that they wanted to visualise (with pencils); lastly, students generate the data visualisation.

After completing the hands-on exercise, students had the ability to independently complete a portion of the data science workflowfrom collecting data to generating visualisations.

3.5.4.2 The Infographic exercise

An infographic is a widely used tool to present statistical information derived using data science methods. We gave students plenty of time to read the infographic (Figure 3.14) with some guidance (the tips are given in the last question) in the teaching material. In daily life we normally do not read all of the content in an infographic; instead, we tend to only read the parts that we are interested in. We followed this principle to teach students how to read infographics. Here, students solved their last question, *(8 min) Question 11: Read the infographic below following the three steps: 1) read the topic; (2) find the panel that interests you; (3) read the panel. If you are stuck, try to find some explanation near your interested panel. What do you learn in this infographic?* The facilitators answered all the questions students asked during this process.

After completing the last hands-on exercise (reading the infographic) with and with the help of the data science knowledge they have gained from our teaching material, students should be able to read most of the information in their lives that is presented in the form of data science.

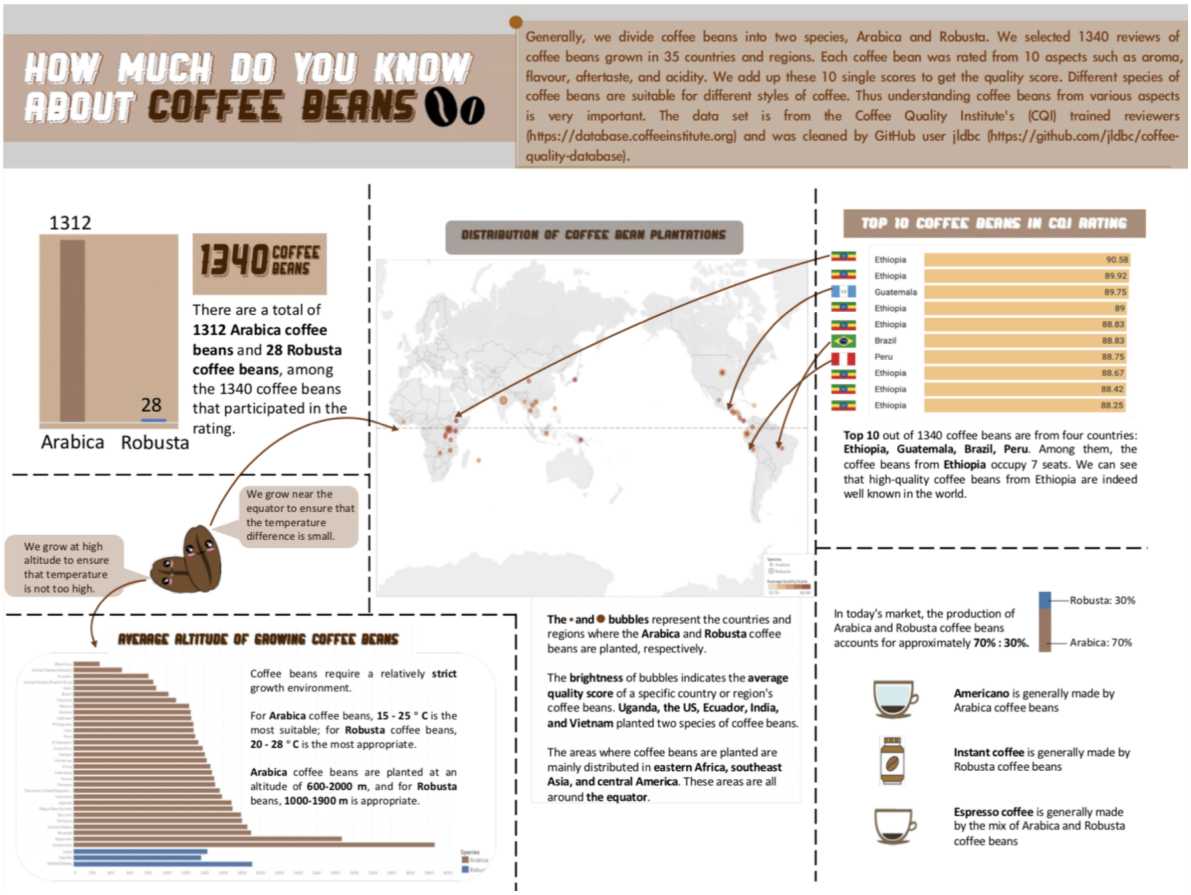


Figure 3.14: The infographic for students to do hands-on exercise.

Chapter 4

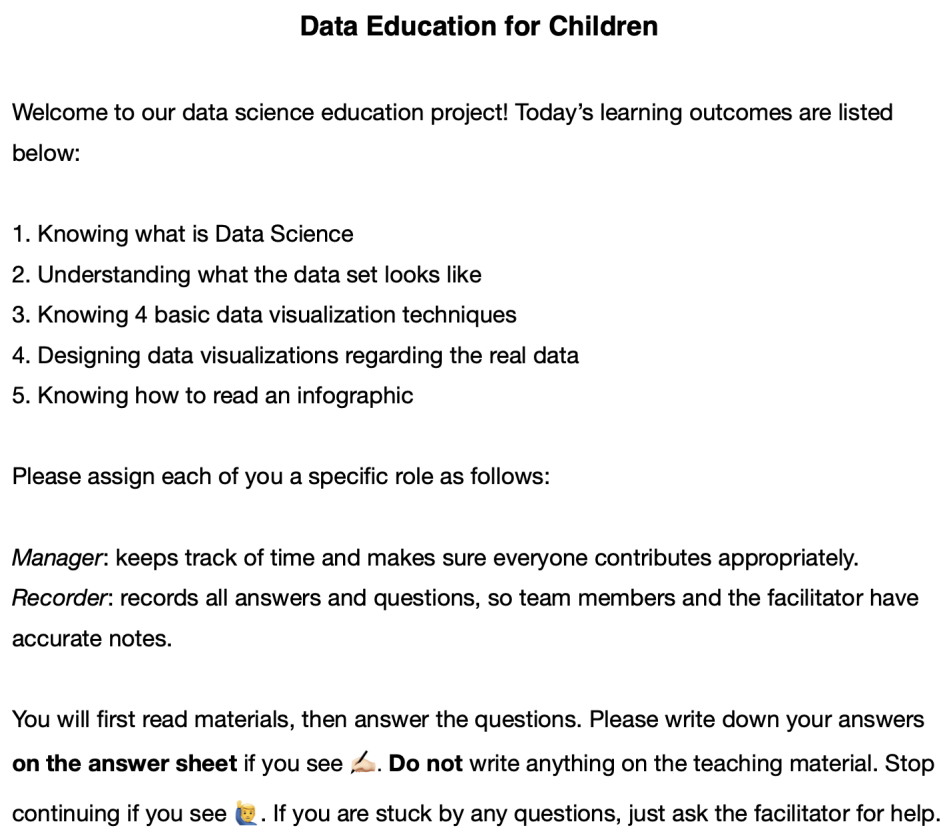
Evaluation

4.1 Evaluated by Exports

Our first evaluation was completed in cooperation with Kate, a high school teacher who taught computer science. She is the Director of Curriculum for the Data Education in Schools project at the University of Edinburgh. Since our teaching material is designed following the POGIL method, we do not have to stand on the platform as the traditional way of teaching. What we need to do is to give the teaching material to the evaluator and answer the questions posed by the evaluator during the process of reading the teaching material.

The first problem arises on the first page. We should specify the way the teaching material is used (similar to the product instruction) and set up little marks to remind the student to respond or take some kind of action. Doing so allows students to fully understand what they should be doing at which part of the process. The improved page is shown in Figure 4.1, and the unmodified version is shown in Figure 3.1. The comparison between the question after and before adding the mark is shown in Figure 4.2.

The second element worth modifying is the first activity of the exploration phase. In the first version, we showed students three examples of data sciences related to everyday life, and set the first question after these examples: *"(3min) Question 1: use your words to conclude what Data Science is."* We hoped that students could use these examples to figure out how data science works and thus learn new information by examining existing data. Returning to the definition of exploration, *students infer certain trends or patterns from the data already provided or the data they collect, and validate the assumed concepts through subsequent questions*, notice "*...students* ***Infer***



Data Education for Children

Welcome to our data science education project! Today's learning outcomes are listed below:

1. Knowing what is Data Science
2. Understanding what the data set looks like
3. Knowing 4 basic data visualization techniques
4. Designing data visualizations regarding the real data
5. Knowing how to read an infographic

Please assign each of you a specific role as follows:

Manager: keeps track of time and makes sure everyone contributes appropriately.
Recorder: records all answers and questions, so team members and the facilitator have accurate notes.

You will first read materials, then answer the questions. Please write down your answers **on the answer sheet** if you see ✍🏽. **Do not** write anything on the teaching material. Stop continuing if you see 🙋‍♂️. If you are stuck by any questions, just ask the facilitator for help.

Figure 4.1: The second version of the first page.

*certain trend...**from the data***". We did provide the reasoning material when designing the teaching material, but we ignored the need for students to reason out the concepts step by step on their own. Kate suggested that we should not directly lead with the question of "what is data science?", but should instead split the question into several small questions to guide students in their thinking. We addressed a question after each example to guide students to think: *(3 min) Question 1: Why these recommendations are not like yours? What data do you think is needed for the recommendation system; (3 min) Question 2: How Do you think scientists predict the weather?; (3 min) Question 3: How do you think Facebook predicts the people you may know?* After finishing these three questions, they are led to the fourth question, *(4 min) Question 4: Use your word to conclude how Data Science works.* The specific examples in the teaching material refer to the Figure 3.2.

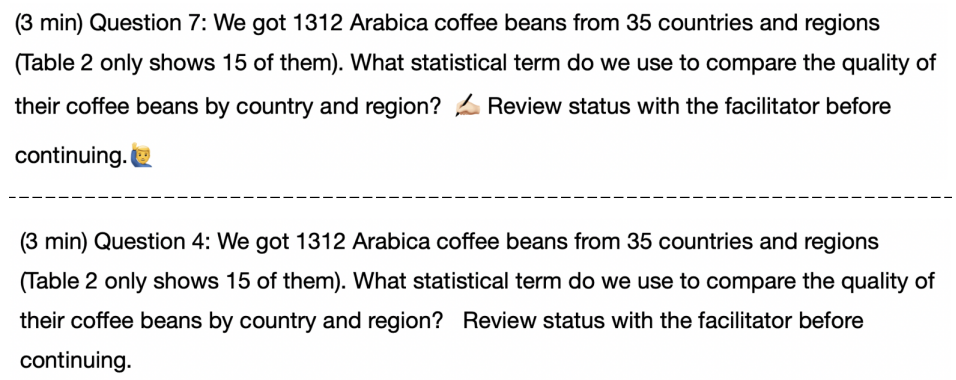


Figure 4.2: The comparison between the question after and before adding the mark. Up: the question after adding the mark; down: the question before adding the mark

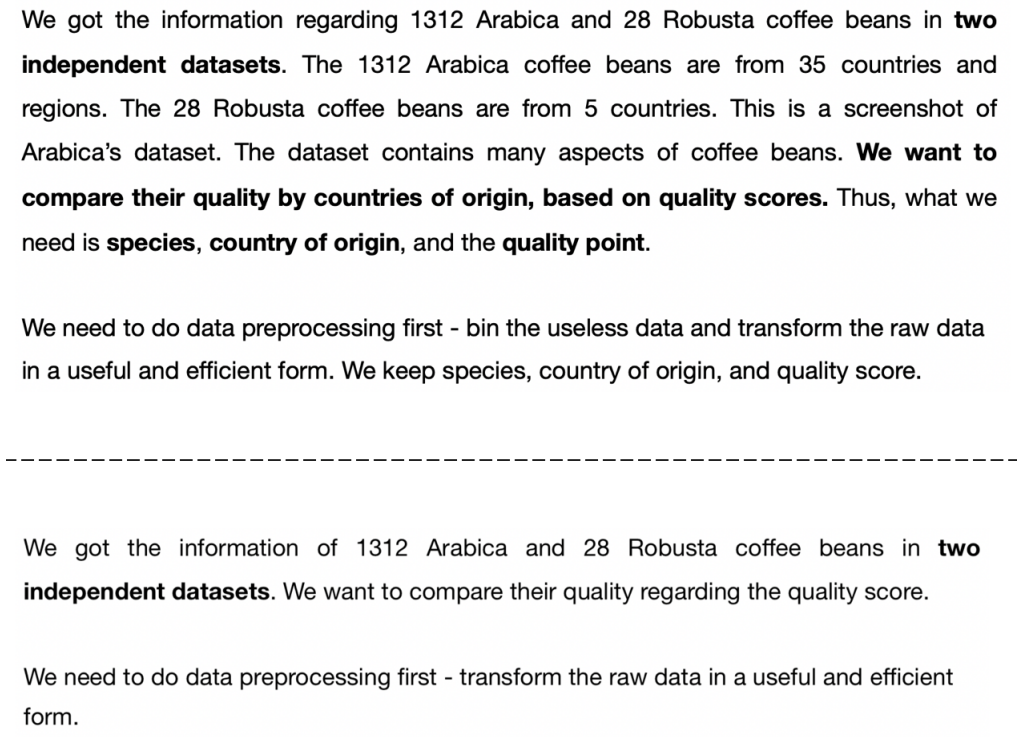


Figure 4.3: The comparison between the two descriptions after and before the modification. Up: the description after the modification; down: the description before the modification

The third problem occurs in the second activity of the exploration phase the coffee bean exercise. We should think about the data set and data science from the perspec-

tive of the students, which means we should not have a prior impression of the data set to be used. Thus, we should be as clear and accurate as possible when describing the data set, and mark the key points we want students to understand. Similarly, we should also describe the principles of data preprocessing, although this process does not require students to have a complete understanding of data processing. The comparison between the two descriptions after and before the modification is shown in Figure 4.3.

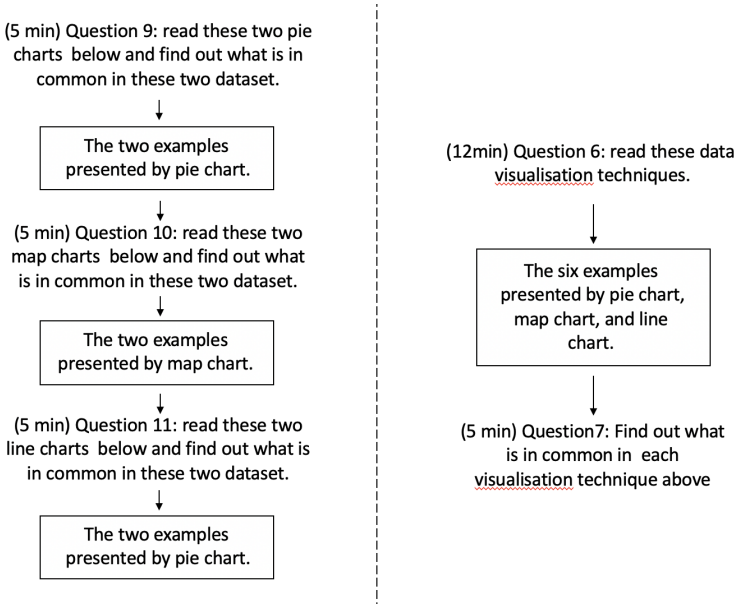


Figure 4.4: The comparison between the two plans of addressing questions after and before the modification. Up: the plan after the modification; down: the plan before the modification

The fourth problem is somewhat similar to the second one. In preparing our concept, our original plan was to let students read the examples represented by the pie chart, map chart, line chart, then summarise the common points in each visualisation technique, and finally infer the characteristics and the scope of application of each visualisation technique. The drawback is that the students see the problem after reading the example, which means they need to re-read the example to find the common points necessary to answer the question, which is a waste of time. Thus our proposed improvement is: let the students understand the task before reading the question by summarising the commonalities of the dataset in the two examples of each visualisation technique. Additionally, we split the question into three smaller questions that would better guide students towards thinking on the right track. The comparison be-

tween our two types of questions after and before the modification is shown in Figure 4.4.

The last problem that Kate pointed out to resolve is the props in the hands-on exercise of the application phase, M&M peanut chocolates. A certain percentage of students are allergic to peanuts. This obstacle, which is outside of the data science and teaching method, is something we never thought about. We replaced the peanut chocolate with a jelly bean, a snack that is safe for all students.

"This teaching material is good. Using POGIL in paper teaching material is easy for the teacher to check the progress of each group and student.", said Kate. Overall, Kate was satisfied with our teaching materials, supporting our belief that adopting POGIL for data science education is groundbreaking work. The application of POGIL allows students to actively inquire the answers to the questions, make progress together in teams, and lets the teacher keep abreast of all her students progress. Presenting examples that are relevant to students' daily life can help them stay engaged with the world of data science. On the other hand, while designing the teaching material, we should follow the perspective of students, describe the reasoning material accurately, and set the questions step by step to guide students to think on the right track.

4.2 Evaluated by Young People

After modifying the teaching material according to Kate's suggestion, we created the second version of the teaching material and accepted the evaluation of two 16-year-old students, P1 and P2. In this evaluation we followed the real classroom environment - students assigned roles themselves before they started exploring the questions and finally collaborating to discuss the answers to each question.

In order to quantify the students' impression of the teaching material accurately, we prepared questionnaires. The questionnaire we used is based on the ARCS motivation model introduced by J.M.Keller [18, 16, 19]. Kellers ARCS Model of motivation can be perceived as a problem-solving approach to learning that instructional designers can use to develop even more engaging learning activities [26]. This model has four factors (attention, relevance, confidence, and satisfaction), and each factor has three sub-categories. Thus, there are twelve lower categories in total. Each item was presented using a five-point scale where answer 6 always corresponded to "agree" and answer 1 to "disagree". We delivered the questionnaire twice first, after students finished the last question in the exploration phase, then again at the end of the lesson. We

compared the results of both to assess which part of the teaching material needs to be improved. The questionnaire and results are shown in Figure 4.5.

Questionnaire		**Results**			
Attention		P1 #1	P1 #2	P2 #1	P2 #2
A1	The data science education gives you a pleasant surprise.	3	3	3	4
A2	You wanted to learn more during the data science education.	3	3	3	3
A3	You could study without getting bored because there were variations in the learning content.	4	3	4	4
Relevance					
R1	You feel that the learning content was familiar.	3	4	5	3
R2	You understand the goal and the importance of the learning.	4	5	4	5
R3	You had chances to select the learning methods that were suitable to you.	3	3	3	4
Confidence					
C1	The goal you should reach is clear.	3	5	3	5
C2	You had an occasion to feel that you had written your answer well.	4	5	4	3
C3	You felt that you wrote your answer well because of your efforts and ability.	4	4	3	4
Satisfaction					
S1	You will have occasions to use your newly acquired knowledge.	4	5	4	5
S2	You were happy when you made a good answer.	4	4	4	5
S3	Everyone in your group contributes appropriately.	5	5	5	4

Figure 4.5: The questionnaire and results of the second evaluation.

The average score of P1 in the first time was 3.67, and the second time was 4.08. The average score of P2 in the first time was 3.75, and the second time was 4.08. Overall, both P1 and P2 gave better feedback during the second evaluation than during the first time, which means that the performance of the *concept invention* and *application* phase is better than that of the *exploration* phase. All the scores of four evaluations are greater than 3, which means that the participants were motivated by the teaching material.

Specifically, in the *attention* factor of the questionnaire, P1 scored 4 for A3 during

the first time and 3 during the second time, which indicates that our teaching material did not make P1 feel like there were variations in the second half. P2 scored A1 from 3 to 4, which shows that P2 gets more and more active when using the teaching material.

In the *relevance* factor, P1's second scores on R1 and R2 are higher than the first; P2's second score on R2 and R3 is higher than the first, and the second score of R1 is lower than the first. We can conclude that the concepts that emerge in the teaching material are generally relevant to what students are learning in schools, however the students have not yet been exposed to all the concepts related to data science. "This is my first contact with the map chart. This visualisation technique has made me feel refreshed. ", P1 said to our facilitators after the lesson.

In the *confidence* factor, we noticed that P2's second scoring of C2 was lower than the first. This shows that as the course progresses, the difficulty of the teaching material gradually increases. P2's recognition of his answers gets lower and lower, and he does not write the answers he thinks are satisfactory. Therefore, we can conclude that in the process of POGIL, the facilitator should always pay attention to the learning status of each student and help catch up the students who are temporarily behind.

In the *satisfaction* factor, the results showed that among all the score changes, the only drop was P2's scoring of S3. As the course progressed, some students could observe that not all team members were working at the same pace and could not contribute appropriately.

Overall, the students' recognition of the teaching materials gets higher and higher as they explore the concepts outlined in this process, indicating that our reasoning material is relevant to students' lives. Additionally, this shows that the questions we set can guide the students to explore the data science step by step, and get the right answer to the questions in the end. However, during the lesson, the facilitator must not only focus on the completion of the groups but also pay close attention to each specific student's learning status and help any students who are temporarily left behind.

Chapter 5

Discussion

Applying POGIL to the field of data science education is a groundbreaking endeavor. We completed a lot of preparatory work in the early stage, including exploring the current development of data science education for young people in worldwide, the characteristics of data science itself, and how to increase engagement of students in data science. After designing the teaching material, we organised two evaluations. The first evaluation was done by a high school computer science teacher, and the second one was by two 16-year-old students. They all thought that our teaching material was interesting and practical for young people.

Our teaching material has generally achieved satisfactory results. The teaching material is rich in hands-on exercises that is relevant to the students' lives and accurately demonstrates basic data science concepts. Applying POGIL to design the teaching material guarantees that the material will be highly interactive and continue to attract students' attention. Future versions of this teaching material will still follow the POGIL way. Returning to our original research question how to set the appropriate questions for students to inquire about the concepts of data science? we should follow a students interests and pose inspiring questions before the more intensive and analytical questions. Only then can we guide students on the right track towards understanding data science.

Looking at the progress participants made in the lesson, the first four steps of the full cycle of data science workflow are not difficult for 16-year-old students to understand. In other words, this teaching material is not a big challenge for 16-year-old students. In future work, we will add content related to supervised learning to the teaching material, such as learning a simplified version of the decision tree, and applying the algorithm to the practical problem. In terms of data sets, we will also extend

to other topics, such as the historical achievements of a team in the Premier League. These kinds of topics may provoke students' interest even more than coffee beans.

The flaw in our teaching material is that when guiding students to explore problems, they always follow the same approach - reading reasoning materials and inquiry questions. This will make the students think that there are not enough variations in the teaching material. In the future, we will set more reasoning methods to enrich the fun of the teaching material. In the evaluation session, our evaluator group consisted of only one high school teacher and two middle school students. This sample size is not sufficient to account for all the problems of the teaching material. In the future, we will involve more people in the evaluation of our teaching material.

Chapter 6

Conclusions

We tried to adopt POGIL to the field of data science education for young people. We applied POGIL to design the teaching material of data science for young people. We first analysed the necessity for young people to learn data science, and then explored the current status of data science education for young people worldwide. Based on the background, we present three principles that should be followed when designing the data science teaching material - stimulating creativity, stimulating interest, and emphasising quality.

After comparing Active Learning and POGIL, we chose to design the teaching material in the POGIL method. Data science includes many diverse aspects, so our teaching material focused specifically on the knowledge level of young people. Our aim is to guide the students to experience the first four steps of the full cycle of data science workflow. When designing the teaching material, we paid special attention that the hands-on exercises should be relevant to the students' lives while still demonstrating data science concepts. After completing the teaching material, we organised two evaluations. The first evaluation was done by a high school computer science teacher, and the second one was by two 16-year-old students.

In conclusion, our teaching material was ultimately successful. We have succeeded in finding a balance between students' natural interest and learning data science concepts. With the development of data science, more and more teaching materials and courses related to this field will appear in the future. We hope that our attempts can bring some inspiration to the development of this industry.

Bibliography

[1] https://github.com/jldbc/coffee-quality-database.

[2] Coffee quality institude. https://database.coffeeinstitute.org.

[3] Exploring computational thinking. https://edu.google.com/resources/programs/exploring-computational-thinking/.

[4] Pogil. https://pogil.org.

[5] About - focus and scope. https://datascience.codata.org/about/, April 2002.

[6] The ieee international conference on data science and advanced analytics. http://www.dsaa.co/, 2013.

[7] International journal on data science and analytics. https://www.springer.com/computer/database+management+2013.

[8] Cheryl L Aasheim, Susan Williams, Paige Rutner, and Adrian Gardiner. Data analytics vs. data science: A study of similarities and differences in undergraduate programs based on course descriptions. *Journal of Information Systems Education*, 26(2):103–115, 2015.

[9] Valerie Barr and Chris Stephenson. Bringing computational thinking to k-12: what is involved and what is the role of the computer science education community? *Inroads*, 2(1):48–54, 2011.

[10] Shengwen Dai. Informatics project proposal: Data education for children, April 2019.

[11] Chamindra de Silva. Humanitarian free and open source software. *Open Source Business Resource*, 12/2010 2010.

[12] Vasant Dhar. Data science and prediction. 2012.

[13] Caitlin Duncan, Tim Bell, and Steve Tanimoto. Should your 8-year-old learn coding? In *Proceedings of the 9th Workshop in Primary and Secondary Computing Education*, pages 60–69. ACM, 2014.

[14] Escoufier et al. *Data Science and its Application*. Tokyo: Academic Press, first edition, 1995.

[15] Ann Bainbridge Frymier and Gary M Shulman. what's in it for me?: Increasing content relevance to enhance students' motivation. *Communication Education*, 44(1):40–50, 1995.

[16] Robert M Gagne, Walter W Wager, Katharine C Golas, John M Keller, and James D Russell. Principles of instructional design. *Performance Improvement*, 44(2):44–46, 2005.

[17] Walter Gander, Antoine Petit, Gérard Berry, Barbara Demo, Jan Vahrenhold, Andrew McGettrick, Roger Boyle, Avi Mendelson, Chris Stephenson, Carlo Ghezzi, et al. Informatics education: Europe cannot afford to miss the boat. *ACM,[online] Available at: http://europe. acm. org/iereport/ie. html*, 2013.

[18] John M Keller. Use of the arcs motivation model in courseware design. *Instructional designs for microcomputer courseware, Lawrence Erlbaum Associates*, pages 401–434, 1987.

[19] John M Keller. *Motivational design for learning and performance: The ARCS model approach*. Springer Science & Business Media, 2009.

[20] Jeff Leek. The key word in "data science" is not data, it is science. https://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/, December 2013.

[21] Steve Lohr. For todays graduate, just one word: Statistics. *The New York Times*, 158(54759):A1, 2009.

[22] James Manyika. Big data: The next frontier for innovation, competition, and productivity. *http://www. mckinsey. com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innova* 2011.

[23] Andrew J Martin and Martin Dowson. Interpersonal relationships, motivation, engagement, and achievement: Yields for theory, current issues, and educational practice. *Review of educational research*, 79(1):327–365, 2009.

[24] Richard Moog and James N. Spencer. *POGIL: An overview*, volume 994, pages 1–13. 09 2008.

[25] Jeanne Ellis Ormrod, Eric M Anderman, and Lynley Hicks Anderman. Educational psychology: Developing learners. 2006.

[26] Christopher Pappas. Instructional design models and theories: Kellers arcs model of motivation. https://elearningindustry.com/arcs-model-of-motivation, May 2015.

[27] Gil Press. Data science: What's the half-life of a buzzword? https://www.forbes.com/sites/gilpress/2013/08/19/data-science-whats-the-half-life-of-a-buzzword/204a40097bfd, August 2013.

[28] Gil Press. A very short history of data science. https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/7134b73d55cf, May 2013.

[29] Robin Roberson. Helping students find relevance. https://www.apa.org/ed/precollege/ptn/2013/09/students-relevance, September 2013.

[30] Richard M Ryan and Edward L Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68, 2000.

[31] D Simberloff, BC Barish, KK Droegemeier, D Etter, N Fedoroff, K Ford, L Lanzerotti, A Leshner, J Lubchenco, M Rossmann, et al. Long-lived digital data collections: enabling research and education in the 21st century. *National Science Foundation*, 2005.

[32] Neil Smith, Clare Sutcliffe, and Linda Sandvik. Code club: bringing programming to uk primary schools through scratch. In *Proceedings of the 45th ACM technical symposium on Computer science education*, pages 517–522. ACM, 2014.

[33] Il-Yeol Song and Yongjun Zhu. Big data and data science: what should we teach? *Expert Systems*, 33(4):364–373, 2016.

[34] Dan Sperber and Deirdre Wilson. Relevance theory. *Handbook of Pragmatics. Oxford: Blackwell*, pages 607–632, 2004.

[35] Shashank Srikant and Varun Aggarwal. Introducing data science to school kids. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, pages 561–566. ACM, 2017.

[36] Rong Tang and Watinee Sae-Lim. Data science programs in us higher education: An exploratory content analysis of program description, curriculum structure, and course focus. *Education for Information*, 32(3):269–290, 2016.

[37] D.J. Patil Thomas H. Davenport. Data scientist: The sexiest job of the 21st century. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century, October 2012.

[38] Jilong Zhang, Anna Fu, Hao Wang, and Shenqing Yin. The development of data science education in china from the lis perspective. *International Journal of Librarianship*, 2:3, 12 2017.

[39] James Ellwood Zull. *The art of changing the brain: Enriching teaching by exploring the biology of learning*. Stylus Publishing, LLC., 2002.

Appendix A

The Teaching Material

Data Education for Children

Welcome to our data science education project! Today's learning outcomes are listed below:

1. Knowing what is Data Science
2. Understanding what the data set looks like
3. Knowing 4 basic data visualization techniques
4. Designing data visualizations regarding the real data
5. Knowing how to read an infographic

Please assign each of you a specific role as follows:

Manager: keeps track of time and makes sure everyone contributes appropriately.
Recorder: records all answers and questions, so team members and the facilitator have accurate notes.

You will first read materials, then answer the questions. Please write down your answers **on the answer sheet** if you see ✍🏻. **Do not** write anything on the teaching material. Stop continuing if you see 🙋🏻‍♂️. If you are stuck by any questions, just ask the facilitator for help.

1

-Exploration-

Let' see three examples regarding Data Science in real life.

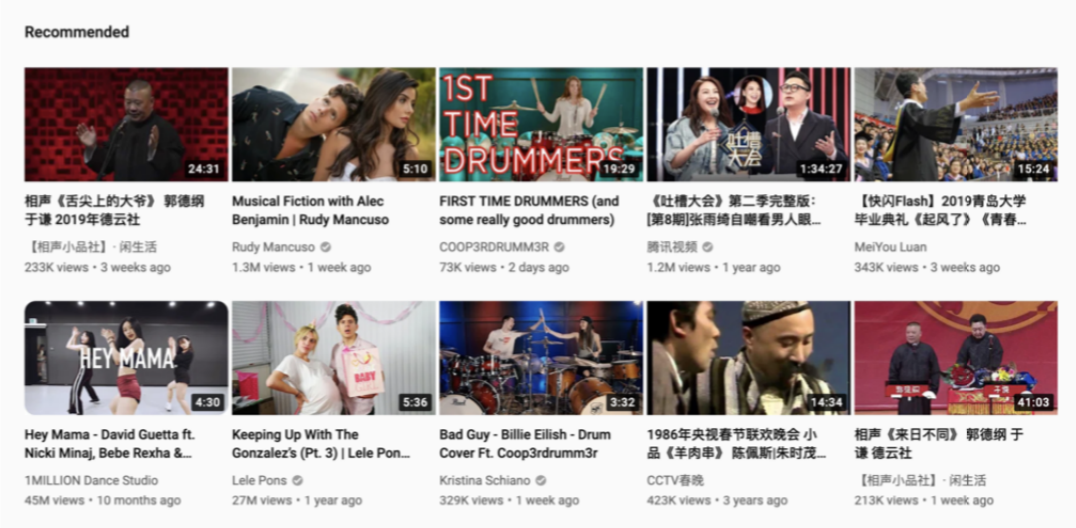


Figure 1: the recommendation system of Youtube

(3 min) Question 1: Why these recommendations are not like yours? What data do you think is needed for the recommendation system? ✍🏻

Figure 2: the weather forecast of Edinburgh.

(3 min) Question 2: How do you think the scientists predict the weather? ✍🏻

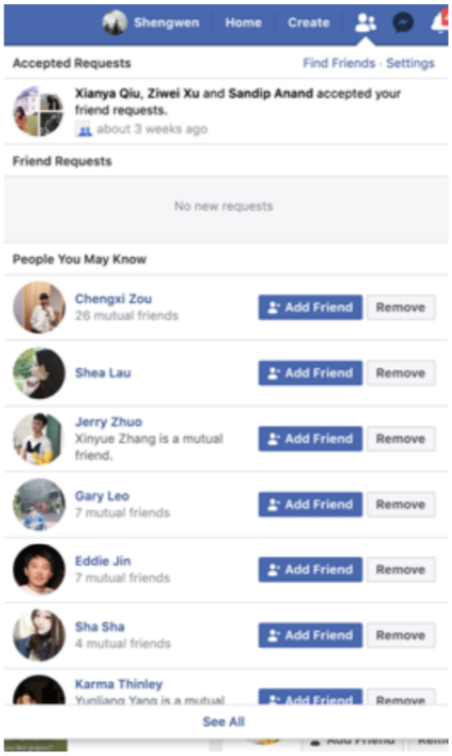


Figure 3: the friends suggestion of Facebook.

(3 min) Question 3: How do you think Facebook predict the people you may know? ✍🏻

(4 min) Question 4: Use your word to conclude how does Data Science work. ✍🏻

Now, let's have a look at today's first dataset.

	quality_score	Species	Owner	Country of Origin	Farm Name	Lot Number	Mill
0	90.58	Arabica	metad plc	Ethiopia	METAD PLC		METAD PLC
1	89.92	Arabica	metad plc	Ethiopia	METAD PLC		METAD PLC
2	89.75	Arabica	Grounds for Health Admin	Guatemala	San Marcos Barrancas "San Cristobal Cuch		
3	89	Arabica	Yidnekachew Dabessa	Ethiopia	Yidnekachew Dabessa Coffee Plantation		Wolensu
4	88.83	Arabica	metad plc	Ethiopia	METAD PLC		METAD PLC
5	88.83	Arabica	Ji-Ae Ahn	Brazil			
6	88.75	Arabica	Hugo Valdivia	Peru	n/a		HVC
7	88.67	Arabica	Ethiopia Commodity Exchange	Ethiopia	Aolme		C.P.W.E
8	88.42	Arabica	Ethiopia Commodity Exchange	Ethiopia	Aolme		C.P.W.E
9	88.25	Arabica	Diamond Enterprise Plc	Ethiopia	Tulla Coffee Farm		Tulla Coffee
10	88.08	Arabica	Mohammed Lalo	Ethiopia	Fahem Coffee Plantation		
11	87.92	Arabica	CQI Q Coffee Sample Representative	United States	El filo		
12	87.92	Arabica	CQI Q Coffee Sample Representative	United States	Los Cedros		
13	87.92	Arabica	Grounds for Health Admin	United States (Hawaii)	Arianna Farms		
14	87.83	Arabica	Ethiopia Commodity Exchange	Ethiopia	Aolme		C.P.W.E
15	87.58	Arabica	CQI Q Coffee Sample Representative	United States	El Águila		

Table 1: The dataset regarding Arabica coffee beans

We got the information regarding 1312 Arabica and 28 Robusta coffee beans in **two independent datasets**. The 1312 Arabica coffee beans are from 35 countries and regions. The 28 Robusta coffee beans are from 5 countries. This is a screenshot of Arabica's dataset. The dataset contains many aspects of coffee beans. **We want to compare their quality by countries of origin, based on quality scores.** Thus, what we need is **species**, **country of origin**, and the **quality point**.

We need to do data preprocessing first - bin the useless data and transform the raw data in a useful and efficient form. We keep species, country of origin, and quality score.

	Species	Country.of.Origin	Total.Cup.Points
1	Arabica	Ethiopia	90.58
2	Arabica	Ethiopia	89.92
3	Arabica	Guatemala	89.75
4	Arabica	Ethiopia	89
5	Arabica	Ethiopia	88.83
6	Arabica	Brazil	88.83
7	Arabica	Peru	88.75
8	Arabica	Ethiopia	88.67
9	Arabica	Ethiopia	88.42
10	Arabica	Ethiopia	88.25
11	Arabica	Ethiopia	88.08
12	Arabica	United States	87.92
13	Arabica	United States	87.92
14	Arabica	United States (Hawaii)	87.92
15	Arabica	Ethiopia	87.83

Table 2: the preprocessed dataset.

(2 min) Question 5: Compare the qualities of the first 5 coffee beans regarding the *Total.Cup.Points* in the table below. ✍🏻

	Species	Country.of.Origin	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	Clean.Cup	Sweetness	Cupper.Points	Total.Cup.Points
1	Arabica	Guatemala	8.42	8.5	8.42	8.42	8.33	8.42	10	10	10	9.25	89.75
2	Arabica	Ethiopia	8.75	8.67	8.5	8.58	8.42	8.42	10	10	10	8.58	89.92
3	Arabica	Ethiopia	8.25	8.5	8.25	8.5	8.42	8.33	10	10	10	8.58	88.83
4	Arabica	Ethiopia	8.67	8.83	8.67	8.75	8.5	8.42	10	10	10	8.75	90.58
5	Arabica	Ethiopia	8.17	8.58	8.42	8.42	8.5	8.25	10	10	10	8.67	89

Table 3: the first 5 coffee beans

(3 min) Question 6: Is there a faster way other than just looking at the figures - what can we transform the figures to? Hint: recall Figure 2. ✍🏻

(3 min) Question 7: We got 1312 Arabica coffee beans from 35 countries and regions (Table 2 only shows 15 of them). What statistical term do we use to compare the quality of their coffee beans by country and region? ✍🏻 Review status with the facilitator before continuing.🙋🏻‍♂️

(3 min) Question 8: The average scores of the coffee beans from Colombia, India, China, Uganda, United States, Ecuador, Ethiopia, Japan are provided below. Complete the bar chart. ✍🏻 Review status with the facilitator before continuing. 🙋🏻‍♂️

	Country of origin	Average Score
1	Colombia	83.11
2	India	76.83
3	China	82.93
4	Uganda	84.05
5	United States	85.98
6	Ecuador	83.83
7	Ethiopia	84.54
8	Japan	84.54

Table 4

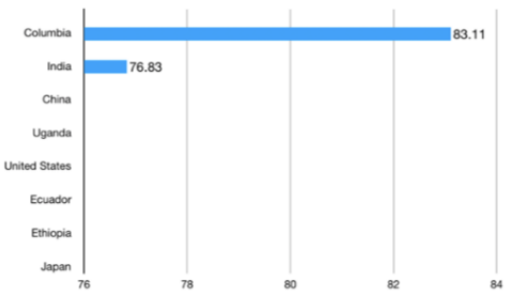


Figure 4

Next, we calculate the average scores of 35 countries and regions and get the entire table and visualisation.

	Country of origin	Average Score (Arabica)
1	Colombia	83.11
2	India	76.83
3	China	82.93
4	Uganda	84.05
5	United States	85.98
6	Ecuador	83.83
7	Ethiopia	84.54
8	Japan	84.54
9	Vietnam	85.75
10	Kenya	84.31
11	Panama	83.71
12	El Salvador	83.05
13	Rwanda	82.83
14	Costa Rica	82.79
15	Thailand	82.57
16	Indonesia	82.57
17	Peru	82.53
18	Brazil	82.41
19	Tanzania	82.37
20	Zambia	81.92
21	Guatemala	81.85
22	Laos	81.83
23	Burundi	81.83
24	United States (Hawaii)	81.82
25	United States (Puerto Rico)	81.73
26	China (Taiwan)	82.00
27	Malawi	81.71
28	Mexico	80.89
29	Philipines	80.83
30	Myanmar	80.75
31	Mauritius	80.50
32	Nicaragua	80.46
33	Honduras	79.36
34	Haiti	77.18
35	Papua New Guinea	85.75

Table 5 **Figure 5**

Do you remember there are **two independent datasets**? The visualisation above is just for Arabica coffee beans. We still need to complement the quality scores of 28 Robusta coffee beans from 5 countries to our data visualisation (this step is to ensure the integrity of the full cycle of data visualisation. You do not need to complement anything).

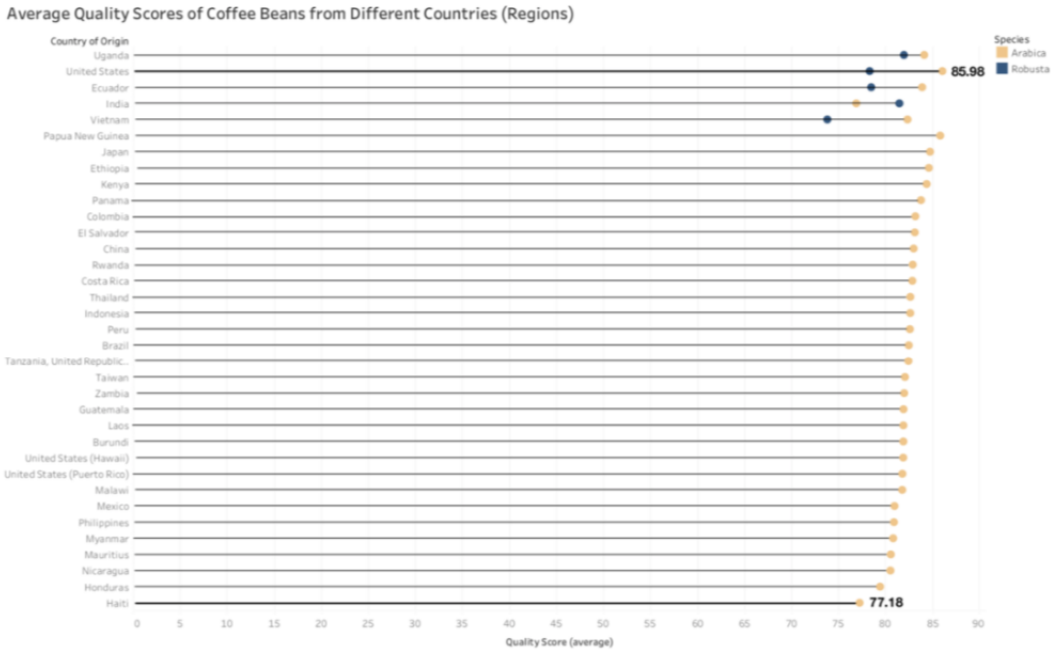


Figure 6: average quality scores of coffee beans from different countries and regions.

You have understood how to do data preprocessing and data visualisation so far. Next section is to learn three widely used visualisation techniques. Review status with the facilitator before continuing. 🙋‍♂️

-Concept Invention-

(5 min) Question 9: Read these two pie charts below and find out what is in common in the two datasets.

Pie Chart #1.

English dialestc	Proportion
American	70.7%
British	15.9%
Canadian	4.9%
Austrilian	4.8%
Other	3.7%

Table 6

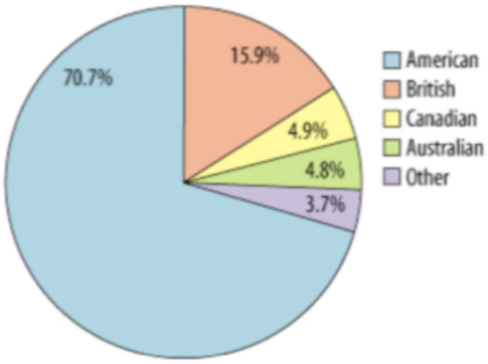


Figure 7

The table and chart show the relative numbers of native English speakers in the major English-speaking countries of the world.

Pie Chart #2.

End Use	proportion
Shower	16.8%
Toilet	26.7%
Leaks	13.7%
Tap	15.7%
Washing Machine	21.7%
Other	5.3%

Table 7

Figure 8

The table and chart show residential end uses of water in 1999. Now please answer question 9. ✍🏻

(5 min) Question 10: Read these two map charts below and find out what is in common in the two datasets.

Map Chart #1.

	Country	Phone Lines
1	Afghanistan	101931.0
2	Albania	235734.0
3	Algeria	3098787.0
4	American Samoa	9900.0
5	Andorra	38237.0
6	Angola	281327.0
7	Antigua and Barbuda	19915.0
8	Argentina	9602605.0
9	Armenia	572772.0
10	Aruba	35000.0
11	Australia	9190000.0
12	Austria	3254700.0
13	Azerbaijan	1795448.0
14	Bahamas	125658.0
15	Bahrain	284684.0
16	Bangladesh	974181.0

Table 8

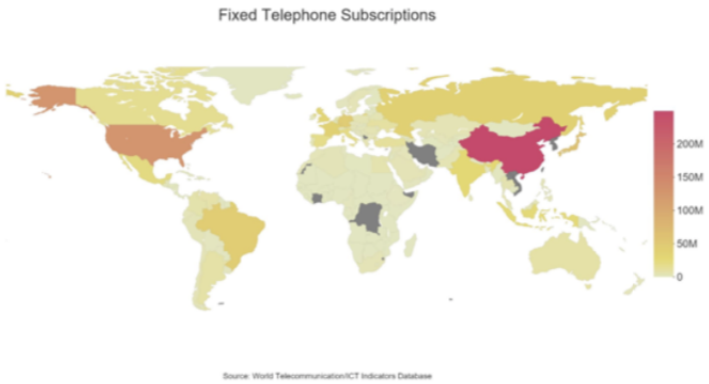


Figure 9

The table and chart show the distribution of fixed telephone subscription in the world.

Map Chart #2.

Name	population
London	7556900
Birmingham	984333
Liverpool	864122
Nottingham	729977
Sheffield	685368
Bristol	617280
Glasgow	591620
Leicester	508916
Edinburgh	464990
Leeds	455123
Cardiff	447287
Manchester	395515
Stoke-on-Trent	372775

Table 9

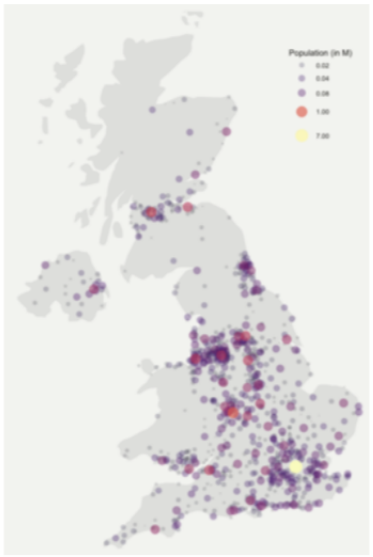


Figure 10

The table and chart show the 1000 biggest cities and towns in the UK by population in 2019. Now please answer question 10. ✍🏻

(5 min) Question 11: Read these two map charts below and find out what is in common in the two datasets.

Line Chart #1.

TIme	Temperature
00:00, Sun 30 Jun	14
01:00, Sun 30 Jun	15
02:00, Sun 30 Jun	15
03:00, Sun 30 Jun	15
04:00, Sun 30 Jun	14
05:00, Sun 30 Jun	14
06:00, Sun 30 Jun	15
07:00, Sun 30 Jun	15
08:00, Sun 30 Jun	16
09:00, Sun 30 Jun	16
10:00, Sun 30 Jun	18
11:00, Sun 30 Jun	16
12:00, Sun 30 Jun	17
13:00, Sun 30 Jun	18

Table 10

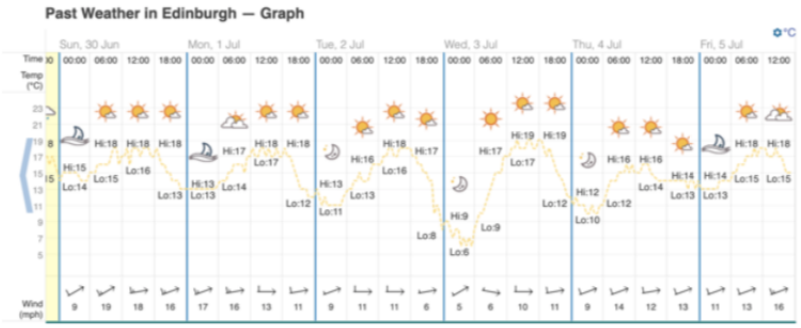


Figure 11

The table and chart show the weather of Edinburgh from 30/06 to 05/07.

Line Chart #2.

Social Media Traffic Referrals (September 2012 – September 2013)

Source	Share of visits September	Share of visits October	Share of visits November	Share of visits December	Share of visits January	Share of visits February	Share of visits March	Share of visits April	Share of visits May	Share of visits June	Share of visits July	Share of visits August	Share of visits September	13-month average	Growth from Sept'12-Sept'13
Facebook	6.53%	6.45%	6.45%	7.76%	7.24%	7.61%	7.56%	8.95%	8.70%	9.34%	8.95%	9.58%	10.37%	8.11%	58.81%
Pinterest	2.21%	2.38%	2.67%	2.84%	3.73%	3.94%	4.09%	3.76%	3.33%	3.38%	3.29%	2.77%	3.68%	3.24%	66.52%
Twitter	0.76%	0.80%	1.01%	1.08%	1.30%	1.39%	1.28%	1.35%	1.35%	1.25%	1.30%	1.20%	1.17%	1.17%	54.12%
StumbleUpon	0.77%	0.58%	0.64%	0.69%	0.65%	0.55%	0.51%	0.65%	0.56%	0.53%	0.44%	0.43%	0.56%	0.58%	-27.47%
Reddit	0.40%	0.40%	0.46%	0.33%	0.37%	0.30%	0.27%	0.28%	0.40%	0.55%	0.45%	0.27%	0.26%	0.36%	-35.16%
YouTube	0.19%	0.21%	0.21%	0.18%	0.18%	0.19%	0.22%	0.23%	0.22%	0.35%	0.46%	0.49%	0.29%	0.26%	52.86%
LinkedIn	0.05%	0.06%	0.06%	0.06%	0.08%	0.08%	0.07%	0.07%	0.08%	0.07%	0.07%	0.06%	0.07%	0.07%	34.51%
Google +	0.04%	0.05%	0.06%	0.05%	0.05%	0.06%	0.06%	0.07%	0.08%	0.08%	0.08%	0.05%	0.04%	0.06%	6.97%

Presented by: Shareaholic

Table 11

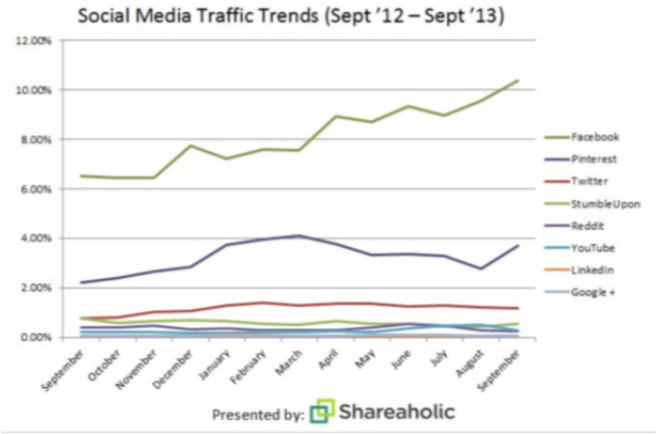


Figure 12

The table and chart show social media traffic trends (Sept'12 - Sept'13). Now please answer question 11. ✍🏻

(4 min) Question 12: Based on the commonalities in each of the visualisation techniques you just find out, conclude in what situations do we use pie chart, map chart, and line chart? ✍🏻 Review status with the facilitator before continuing. 🙋‍♂️

(5 min) Question 13: Based on the answer of question 12, what data visualization other than bar chart can we use to compare the quality of coffee beans by different countries and regions? Why? Sketch the visualisation to verify your assumption. ✍🏻 Review status with the facilitator before continuing. 🙋‍♂️

We have learned bar chart, pie chart, line chart, and map chart so far. In next section we will experience the full cycle of data visualisation - data collection, data entry, data preprocessing, data visualisation, using the the real data you create.

-Application-

Let's finish a small task with jelly beans.

(2 min) Question 14: Open jelly beans and taste some of them.

(12 min) Question 15: In the remaining chocolate, select one or more features (ask the facilitator if you are not sure about what 'feature' is). Then generate your dataset (in the table form) and express it in a visualization way that you think is appropriate. ✍🏻

	Feature 1	Feature 2	Feature 3	...
1				
2				
3				
4				
5				
6				
...				

Table 12: the template of the table.

(4 min) Question 16: do you find something interesting in your visualisation? Review status with the facilitator before continuing. 💪🏻

(8 min) Question 17: Read the infographic below following the three steps: 1) read the topic; (2) find the panel that interests you; (3) read the panel. If you are stuck, try to find some explanation near your interested panel. Is there something interesting in this infographic? ✍

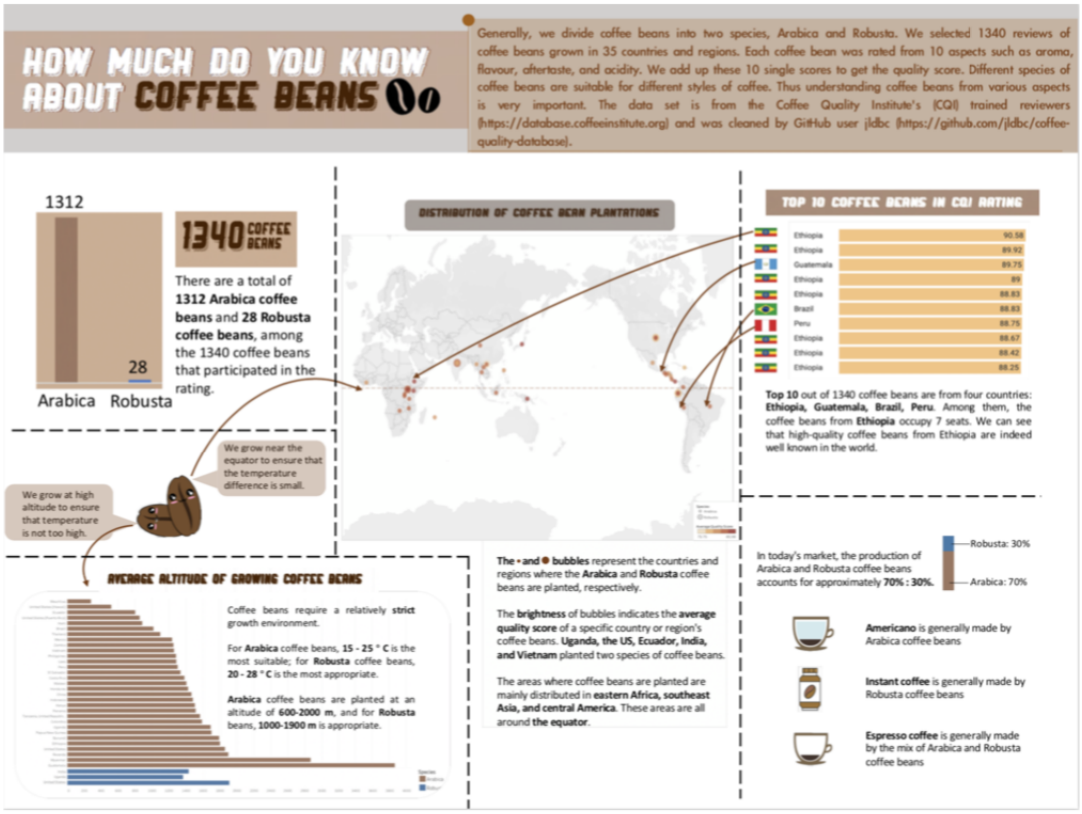


Figure 13: infographic about coffee beans

So far, we have learned what is data science, and we have experienced the full cycle of data visualisation. Also, we understand how to read an infographic now. With the concepts we learned today, we can generate data visualisation using real world data, and read all the hot news related to data science!

Appendix B

The Consent Form

THE UNIVERSITY *of* EDINBURGH

Data Education in School Educational Material Consultation-Information Sheet

You are invited to evaluate a set of educational resources designed to teach children and young people about topics in data science. The materials have been created by MSc students in Informatics at the university of Edinburgh.

During the session you will work through the educational materials which the students have created and discuss your opinions about their suitability for young learners with the MSc students. If you decide that you no long wish to participate, you may withdraw at any point. The students will make notes about your comments and suggestions. Your real name will not be used in analysis or write-up of the research. It will be placed by a participant ID number. The students will type up their notes and save them to secure university filespace and deleted after 3 years.

For general information about how University of Edinburgh uses data go to:
https://www.ed.ac.uk/records-management/privacy-notice-research

For further information about this event, please contact:

Judy Robertson

Professor of Digital Learning

Moray House School of Education

University of Edinburgh

Judy.Robertson@ed.ac.uk

0131 651 6249

THE UNIVERSITY *of* EDINBURGH

Please return this slip to the Judy Robertson to indicate whether your child can take part.

I agree to take part in the Data Science trip.	Yes/No
I agree to be photographed and filmed during the Data Science trip.	Yes/No

Participant's name:

Signature:

Date: